MASTER'S DEGREE THESIS

MSc in Statistics and Operations Research

Title: Joint Modeling Techniques for Analyzing Survival and Longitudinal Data with Applications to the European Randomized Screening for Prostate Cancer (ERSPC).

Author: Xavier Piulachs Lozada-Benavente

Advisor: Carles Serrat i Piè

Co-advisor: Montserrat Rué i Monné

Department: Applied Mathematics I, UPC - BarcelonaTECH

Date: October 2013

Academic year: 2013-2014

Facultat de Matemàtiques i Estadística

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Joint Modeling Techniques for Analyzing Survival and Longitudinal Data with Applications to the European Randomized Screening for Prostate Cancer (ERSPC)

Author: Xavier Piulachs Lozada-Benavente

Advisor: Carles Serrat i Piè
Department of Applied Mathematics I, UPC-BarcelonaTECH

Co-Advisor: Montserrat Rué i Monné
Biomedical Research Institute of Lleida, UdL

October 2013

AGRAÏMENTS

Vull agrair tota l'ajuda i dedicació rebuda dels meus directors de Treball de Fi de Màster, el **Carles Serrat** i la **Montserrat Rué**. M'han acompanyat i orientat en tot moment durant la realització d'aquest treball, i per a ells són les meves primeres paraules de reconeixement.

D'altra banda, vull donar les gràcies de forma especial a la directora del Máster en Estadística i Investigació Operativa, la **Guadalupe Gómez** per donar-me l'oportunitat d'accedir al Màster, i perquè en moments d'incertesa sempre ha tingut paraules de claredat i d'ànims. Igualment, un record molt especial cap a dos professors del màster que he pogut tractat molt, i que sempre m'han animat a lluitar pels meus objectius: La **Monica Becue** i l'**Àlex Sànchez**.

Moltes gràcies també a la meva tutora dins del màster, la **Carme Ruiz de Villa**, pel seu assessorament i consells. D'igual forma, als membres del grup de recerca GRASS per les seves observacions sobre punts a millorar treball, amb una sentida gratitud cap al **Klaus Langohr**, qui sempre ha tingut una estona per a respondre'm dubtes conceptuals i de programació.

Aquests darrers dies de redacció han estat dificils per qüestions de temps, i vull agrair als membres del grup de recerca RISK, i en particular a la **Montserrat Guillén** i al **Ramon Alemany**, la flexibilitat que des de l'inici m'han donat per a compatibilitzar les meves obligacions laborals amb poder concloure la redacció d'aquest treball.

Per a finalitzar, dono les gràcies al **Marcos Luján**, metge responsable de la gestió de les dades espanyoles de l'estudi europeu ERSPC, i que han estat la base de partida per a l'elaboració del present treball.

Mònica, gràcies per estar al meu costat en tot moment.

ABSTRACT

Currently, prostate cancer is the second most frequently diagnosed cancer in European Western countries, specially in Nothern Europe. Compared to other parts of the world, Europe is among the regions with highest incidence. In particular, prostate cancer is a relevant health problem in Spain, where the mean incidence was 57.2/100 000 men-year in 2008 (Ferlay et al., 2010)

Increasing age and some genetic and ethnic risk factors have been identified as causes of prostate cancer development. In this regard, the prostate specific antigen (*PSA*) has become the main marker used for the early detection of prostate cancer, although the precise contribution of this variable to predict disease incidence is not yet very known. Consequently, there is a sense important clinical need for achieving a better knowledge of the marker, so that the increasing number of diagnosed men can be appropriately managed in order to avoid overdiagnosis, i.e., cases in which prostate cancer would have never been detected in the subject's lifetime without screening and would have never progressed to lethal disease. To solve this uncertainty, the *European Randomized Screening for Prostate Cancer* study, ERSPC, was initiated in the early 90's of the last century in order to evaluate the effect of the *PSA* testing on events rates from prostate cancer.

This work analyzes with the data of the Spanish branch of the ERSPC study, where among other longitudinal covariates, *PSA* repeated measurements were taken on the recruited men over time. Moreover, prostate cancer incidence was recorded for each of these subjects. Thus, the aim of this work is to assess the association between the subject-specific profile evolution with the prostate cancer risk, performing a joint analysis of longitudinal and time-to-event data. In the joint modeling approach, the longitudinal covariates are assumed to be of parametric form with random effects (Laird and Ware, 1982), such as a linear mixed effects model, while the Cox proportional hazards model (Cox, 1972) is used to describe the survival information.

The proposed joint modeling techniques were applied to the motivating `PCa Dataset`. The joint model's fixed parameters were estimated with the maximum likelihood method using the Expectation-Maximization algorithm, and for the random effects an empirical Bayes approach was employed. Therefore, a joint model was obtaining connecting the longitudinal and survival processes. The main goal was to measure the association between the true longitudinal *PSA* response (i.e., without measurement error) and the risk for prostate cancer diagnosis, while accounting the special features of each subject. The joint model's assumptions were validated by residual plots, and in the last part of the study the joint modeling results were summarized and discussed, as well as considerations on further areas of research.

All the analysis included in this work have been implemented in the `R` software environment for statistical computing and graphics, using (among others) the following available packages: `nlme` (Pinheiro et al.), `survival` (Therneau, 2012) and `JM` (Rizopoulos, 2010).

RESUM

En l'actualitat, el càncer de pròstata és el segon tipus de càncer més diagnosticat als països d'Europa occidental, amb un gran nombre d'afectats en aquells països europeus més septentrionals. En comparació amb altres parts del món, Europa es troba entre les regions amb un major nivell de repercussió de la malaltia. Particularment, el càncer de pròstata constitueix un greu problema de salut pública a Espanya, on la incidència mitjana va ser de 57.2/100 000 homes-any durant 2008 (Ferlay et al., 2010).

L'augment de l'edat i alguns factors de risc de tipus genètic i ètnic han estat tradicionalment identificats com a causes que afavoreixen el desenvolupament de càncer de pròstata. En aquest punt, el denominat antígen específic prostàtic (*PSA*) s'ha convertit en el principal biomarcador utilitzat per a la detecció precoç de la malaltia, encara que l'exacta incidència d'aquesta variable eś avui en dia molt desconeguda. Existeix per tant una necessitat clínica d'assolir un millor coneixement d'aquest biomarcador, de manera que el cada cop més elevat nombre d'homes diagnosticats pugui ser adeqüadament tractat i s'evitin casos de sobrediagnòstic, és a dir, aquells casos en què el càncer de pròstata mai hauria estat detectat fora d'un procés de cribatge ni tampoc hauria tingut una progressió letal per a la salut del subjecte. Per a millorar el coneixement sobre aquesta incertesa, es va iniciar a principis dels anys 90 del segle passat l'estudi *European Randomized Screening for Prostate Cancer*, ERSPC, evaluant l'efecte de la prova de nivell de *PSA* en la detecció de nous casos de càncer de pròstata.

El present treball analitza les dades corresponents a la secció espanyola de l'estudi ERSPC, on entre d'altres variables es van obtenir en els subjectes de l'estudi mesures repetides del valor de la *PSA* al llarg del temps. D'altra banda, també es va recollir la incidència del càncer de pròstata entre cadascun dels subjectes de l'estudi. D'aquesta forma, l'objectiu d'aquest treball ha estat el poder avaluar si hi ha relaci entre l'evolució particular d'un determinat subjecte amb el risc d'experimentar la malaltia, realitzant per a això una modelització conjunta de les dades longitudinals i de supervivència. Sota una aproximació de tipus *joint modeling*, les covariants longitudinals son tractades de forma paramètrica amb la incorporació d'efectes aleatoris (Laird and Ware, 1982), mentre que les dades de supervivència acostumen a la literatura a ser tractades amb el model de riscos proporcionals desenvolupat per Cox (Cox, 1972).

Les tècniques de *joint modeling* van ser aplicades al conjunt de dades de l'estudi, `PCa Dataset`. Els paràmetres del model es van estimar mitjançant el mètode de màxima versemblança amb l'algoritme *Expectation-Maximization*, i per a predir els efectes aletoris es va utilitzar una aproximació del tipus *Empirical Bayes*. Així, es va poder obtenir un model conjunt que connectés els processos longitudinal i de supervivència. S'ha pogut doncs mesurar el grau d'asscociació entre la resposta longitudinal real (és a dir, sense estar sotmesa a error de mesura) i el risc de diagnòstic de càncer de pròstata, tot considerant les particularitats de cada individu. Les hipòtesis han estat validades per gràfics de residus, i es resumeixen els resultats de la modelització conjunta, així com també es presenten una sèrie de consideracions sobre futures àrees de recerca en el tema.

Totes les análisis incloses en aquest treball han estat implementats amb el programa estadístic de lliure accés `R`, utilitzat per a la modelització de dades i la realització de gràfics. Entre d'altres paquets del programa, s'han fet servir els següents: `nlme` (Pinheiro et al.), `survival` (Therneau, 2012) i `JM` (Rizopoulos, 2010).

RESUMEN

En la actualidad, el cáncer de próstata es el segundo tipo de cáncer más diagnosticado en los países de Europa occidental, con un gran número de afectados en aquellos países más septentrionales. En comparación con otras partes del mundo, Europa se halla entre las regiones con una mayor nivel de repercusión de la enfermedad. En concreto, el cáncer de próstata constituye una grave problema de salud pública en España, donde la incidencia media fue de 57.2/100 000 hombres-año durante 2008 (Ferlay et al., 2010).

El aumento de la edad y algunos factores de riesgo de tipo tanto genético como étnico han sido tradicionalmente identificados como causas que favorecen el desarrollo del cáncer de próstata. En este punto, el denominado antígeno específico prostático (*PSA*) se ha convertido en el principal biomarcador utilizado para la detección precoz de la enfermedad, si bien que la exacta incidencia de esta variable es a día de hoy muy desconocida. Existe por tanto una necesidad clínica imperiosa de tener un mejor conocimiento de este marcador biológico, de manera que el cada vez mayor número de hombres diagnosticados pueda ser adecuadamente tratado y se eviten casos de sobrediagnóstico, es decir, casos en los que el cáncer de próstata nunca se habría detectado fuera de un proceso de cribado, ni tampoco habría tenido una progresión de consecuencias letales para la salud del sujeto. Para solventar esta incertidumbre, a principios de los años 90 del siglo pasado se puso en marcha el estudio *European Randomized Screening for Prostate Cancer*, ERSPC, evaluando el efecto de la prueba de nivel de *PSA* en la detección de nuevos casos de cáncer de próstata.

El trabajo analiza los datos de la sección española del estudio ERSPC, donde, entre otras variables, se recogió en los diferentes sujetos una serie de medidas repetidas del valor de la *PSA* a lo largo del tiempo. También se registró la incidencia del cáncer de próstata en cada uno los sujetos. Así, el objetivo del trabajo es relacionar la evolución particular de un determinado individuo con su riesgo de experimentar la enfermedad, modelizando de forma conjunta los datos longitudinales y de supervivencia. Bajo una aproximación de tipo *joint modeling*, las covariantes longitudinales son tratadas de forma paramétrica con la incorporación de efectos aleatorios (Laird and Ware, 1982), mientras que los datos de supervivencia acostumbran a ser tratados con el modelo de riesgos proporcionales desarrollado por Cox (Cox, 1972).

Les técnicas de *joint modeling* se aplicaron sobre el conjunto de sujetos del estudio, `PCa Dataset`. Los parámetros del modelo se estimaron mediante el método de máxima verosimilitud con el algoritmo *Expectation-Maximization*, y para predecir los efectos aleatorios se empleó una aproximación del tipo *Empirical Bayes*. Así, se ha obtenido un modelo que conecta los procesos longitudinal y de supervivencia, permitiendo medir el grado de asociación entre la respuesta longitudinal real (es decir, sin error de medida) y el riesgo de diagnóstico de cáncer de próstata, considerando las particularidades de cada individuo. Las hipótesis fueron validadas por gráficos de residuos, y se resumen los resultados de la modelización conjunta, así como también se presentan una serie de consideraciones sobre futuras líneas de investigación en el tema.

Todos los análisis incluidos en este trabajo han sido implementados con el programa estadístico de libre acceso `R`, utilitzado para la modelización de datos y la realización de gráficos. Entre otros paquetes del programa, se han utilizado los sigüientes: `nlme` (Pinheiro et al.), `survival` (Therneau, 2012) y `JM` (Rizopoulos, 2010).

CONTENTS

List of Tables

List of Figures

CHAPTER 1
INTRODUCTION AND GOALS

1.1 Motivation

It has become increasingly common in survival studies to record the values of key longitudinal covariates until the occurrence of some particular event of interest $\mathcal{E}$ on a subject. For example, in many medical studies, it is usually collected patients' information repeatedly over time, being of interest the time to recovery or recurrence of a disease. Some longitudinal covariates (e.g., biomarkers) may involve measurement error due to biological processes inherent to the subject or may be affected by non-random dropout. All these difficulties can be circumvented by including random effects in the longitudinal covariates, for example through a linear mixed-effects model, and modeling the longitudinal and time-to-event data jointly rather than separately. Thus, the so-called *Joint Models* consider these two data types together into a single statistical model, so that one can assess the association between the longitudinal measurements and the time to the event of interest under study.

If the analysis focuses on longitudinal data, informative missing values need to be addressed since dropouts are very common in longitudinal studies, whereas if the analysis focuses on survival data, time-dependent covariates must be incorporated as the times to event may be associated with the covariate trajectories. Sometimes, the main interest may lie in the association between the longitudinal process and survival process.

In general, a joint modeling approach is required in mainly three situations:

1. The interest is on the event outcome and one wishes to account for the effect of the longitudinal outcome as a time-dependent covariate. Taditional approaches for analyzing time-to-event data (such as the partial likelihood for the Cox proportional hazards models) are not applicable if the time-dependent covariates are internal or endogenous.

2. The interest is on the longitudinal outcome. In this case the occurrence of events causes dropout since no longitudinal measurements are available at and after the event time. When this dropout is non at random (i.e., the probability of dropout depends on unobserved longitudinal responses), then bias may arise from an analysis that ignores the dropout process.

3. The main interest lies in the association between the true longitudinal profile of the subject and the survival time. To solve this question and obtain valid inferences, the longitudinal and dropout process must be jointly modeled.

1.2 Scope and aims of the work

This study is motivated by the *European Randomized Screening for Prostate Cancer* study, ERSPC, which is the world's largest randomized prostate cancer screening study, where the objective was to investigate whether early detection and treatment of prostate cancer might reduce disease-specific mortality and also help to identify men at risk. In particular, the motivating dataset corresponds to the screening arm of the ERSPC Spanish branch. The main aim of this research work is to show

how joint models can be used to incorporate past and current true *PSA* levels into a predictive model of prostate cancer with the objective of refining the estimation of the time until prostate cancer diagnosis event. To achieve this, the specific goals are the following:

1. To analyze the dataset derived from the Spanish section of ERSPC study, considering the provided information from a dual approach: longitudinal *PSA* data analysis and survival time until the occurrence of the prostate cancer occurrence.
2. To understand all the mathematic methodology related to the longitudinal and survival models, laying the principles of a jointly modelization for both type of models through the so-called *joint models*. In particular, the present study takes as a main reference the progress made in this field by professor Rizopoulos.
3. To apply the joint modeling techniques to a particular case study, such as the referred to the Spanish section from ERSPC project. As a main result, the joint model implementation allows to link a particular subject-specific trend over time to the probability of event diagnosis.
4. To assess the joint goodness-of-fit of the estimated joint model and illustrate how time to prostate cancer diagnosis can be predicted for each individual. This goal enables professionals to make a personalized diagnosis of how disease will evolve, that means, make predictions taking into account the specific characteristics of each individual.

The rest of the paper is organized as follows. Chapter 2 presents a description of the motivating dataset, the Spanish ERSPC data. In Chapter 3, the joint model methodology for longitudinal measurements and time-to-event data is introduced. Chapter 4 shows the results of joint modeling approach for longitudinal *PSA* responses and survival time until prostate cancer diagnosis, including the joint model diagnostics and dynamic survival predictions. Finally, Chapter 5 contains a discussion focused on the impact of longitudinal *PSA* values and baseline covariates on prostate cancer risk, and also the potential areas of future research.

CHAPTER 2

PROSTATE CANCER AND THE ERSPC PROJECT

2.1 Problematic associated to prostate cancer

2.1.1 Prostate cancer incidence and risk factors

Prostate cancer is a form of cancer that develops in the prostate, a gland in the male reproductive system. Most prostate cancers are slow growing and they do not dramatically worsen over time, so an "active surveillance" with biopsy tests may be a good option for these cases. However, there are cases of aggressive prostate cancers which need a more intensive treatment that includes tumor removal through surgery or radiation therapy. The cancer cells may metastasize (spread) from the prostate to other parts of the body, particularly the bones and lymph nodes. Prostate cancer may cause pain, difficulty in urinating or problems during sexual intercourse. Other symptoms can be potentially developed during later stages of the disease.

Rates of detection of prostate cancer vary widely across the world (Figure 2.1), with South and East Asia detecting less frequently than in Europe, and especially in the United States (Swerdlow et al., 2008). Prostate cancer tends to develop in men over the age of fifty, and it is the sixth leading cause of cancer death among men worldwide with an estimated 899 100 new cases and 258 100 deaths in 2008. In particular, 370 700 new cases were detected in Europe, and 89 600 men died of the disease (Ferlay et al., 2010). However, many men with prostate cancer never have symptoms, undergo no therapy, and eventually die of other unrelated causes. Many factors, including genetics and diet, have been implicated in the development of prostate cancer.

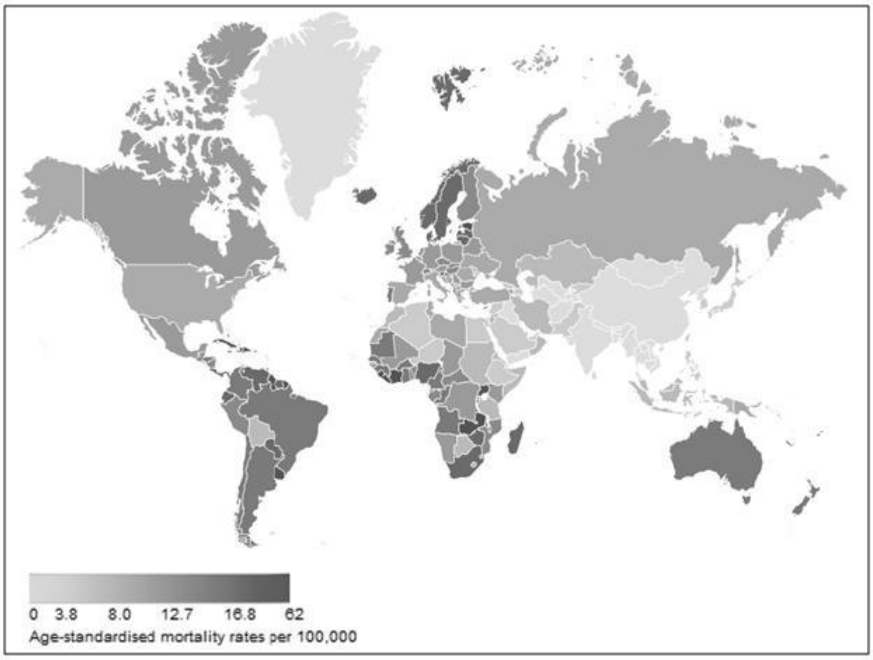

Figure 2.1. Age-standardized prostate cancer mortality rates / 100 000 globally in 2008 (Swerdlow et al., 2008).

In terms of disease research, most studies (Brandt et al., 2003; Schröder et al., 2012) predict the presence of prostate cancer together based on three types of test: the recorded level of prostate-specific antigen (*tPSA* or simply *PSA*), the value of prostate-specific antigen ratio (*PSARAT*) and by physical examination of the subject. Below, we describe briefly the main diagnostic tests that can lead to the diagnosis of the disease:

1. Total prostate specific antigen, *PSA*

   The prostate-specific antigen is a protein found in the blood, which is produced almost exclusively by epithelial prostatic cells, and it is usually measured in nanograms per milliliter of blood (ng/ml). It represents a useful biomarker which is used for the early detection of prostate cancer. In this regard, a high blood *PSA* level tends to be associated with an increased risk of having prostate cancer. However, an elevated *PSA* level may also be due to other causes: a) an increase in normal prostate glands like in benign prostate hyperplasia (BPH) or b) an increased leakage of *PSA* into the bloodstream due to infectious processes or obstruction. Consequently, an elevated *PSA* score often does not necessarily indicate the presence of disease. Moreover, a low *PSA* level does not exclude the possibility of prostate cancer: even in very low *PSA* ranges, a considerable prostate cancer detection rate has been described.

2. Prostate specific antigen ratio, *PSARAT*

   It is defined as the relationship between free *PSA* (*FPSA*, ng/ml) and total *PSA* (ng/ml): $PSARAT = FPSA/PSA$. The free *PSA* is the small *PSA* amount in the blood which is not protein bound. The importance of *PSA* ratio is that the risk of cancer diagnosis increases significantly if the free to total ratio is less than a certain percentage (typically $15\%-25\%$).

3. Physical examination of the subject

   - Digital Rectal examination, *DRE*: A digital examination is a simple technique where the doctor inserts a lubricated finger through the subjects rectum to feel for suspicious areas of the prostate gland. This allows to check those lumpy or abnormal areas on the prostate. The exam is only limited to detecting lesions protruding the surface of the prostate. A digital rectal exam is done for men as part of a complete physical examination to check the prostate gland size.

   - Transrectal ultrasound test, *TRUS*: It is nowadays routinely performed when assessing the status of the prostate. With an ultrasound probe a urologist can visualize the prostate through the rectum wall, estimate the size of the prostate, assess the homogeneity of the parenchyma, and observe possible cancer growth through the capsule of the prostate. The benefits of *TRUS* are minimal invasiveness, relatively low cost and the fact that no ionizing radiation is needed. The sensitivity of conventional *TRUS* has been estimated at $39\%-75\%$ and specificity at $40\%-82\%$ (Heijmink et al., 2011).

Positive results in all the tests above described are taken as signals of a possible presence of prostate cancer, but in none of them can be taken as conclusive. The only reliable way to diagnose a prostate cancer necessarily involves performing a biopsy test. From the biopsy result, the doctor can determine whether there is cancer present, its aggressiveness and the likelihood of spreading. Thus, **only a biopsy can definitively diagnose prostate cancer**.

If the initial tests (*PSA* level, rectal examination or ultrasound test) show that there is a possibility of prostate cancer, a man may be offered a biopsy, in which several samples of tissue (usually around 10) are taken from the prostate to be looked at under a microscope. Most centers used a *PSA* cut-off value of 3.0 ng per milliliter as an indication for biopsy. Choosing a threshold serum level in a screening setting will thus be a trade off between sensitivity and specificity.

2.1.2 Prostate cancer screening

With an estimated 258 100 deaths, worldwide in 2008, prostate cancer is the sixth leading cause of death from cancer in men, representing the 6.1% of the total (Ferlay et al., 2010). Comprehensibly, with so many peoples' lives affected by this disease, early detection has become the aim of many center researchers, although it has become extensively discussed. Screening for disease in asymptomatic individuals has an intuitive interest. If all cancers have a progression toward mortality, then the best strategy seems to stop the progress of the disease as early as possible.

Screening, in medicine, is a strategy used in a population to identify a specific disease in those individuals without signs or symptoms. This can include individuals with pre-symptomatic or unrecognized symptomatic disease. As such, screening tests are somewhat unique in that they are performed on persons apparently in good health. In particular, many studies apply the called "mass screening", which means the screening of a whole population or a subgroup, so that a particular test is offered to each individual, irrespective of the risk status of the individual. Examples of diseases in which mass screening techniques are applied can be breast cancer or cervix cancer.

In the case of prostate cancer, there are different tests used, but the main is the *PSA* level test (in recent years accompanied by a *PSARAT* measurement), which defines the concentration of this molecule in the blood. Other alternative tests are basically based on a physical examination: *DRE* and *TRUS* tests.

The main sources of information that currently can be found about prostate cancer screening are the reports on the two world's important studies on the subject: the *European Randomized Screening for Prostate Cancer* study (Schröder et al., 2012), hereafter the ERSPC, involving 182 000 men aged 45 to 75 years from 8 participating countries, and the *American Prostate, Lung, Colorectal and Ovary Cancer Screening Trial* (Andriole et al., 2009), PLCO, with 76 000 men aged 55–74 from 10 USA participating centers. The characteristics of ERSPC and PLCO studies were similar if not identical. Thus, both of the studies had the objective of evaluating the effect of *PSA* screening on death rates from prostate cancer.

However, much of the controversy regarding the efficacy of prostate cancer screening was originated due to the divergence between the conclusions of these two studies. The European study showed a reduced death rate ratio in the screening group compared to the control of 0.8 (95% confidence interval 0.65−0.98), i.e., it concluded that screening reduced mortality by 20%. In contrast, the American data showed no statistically significant difference in death rates between the screened and control group, but there were more prostate cancer deaths in the screened group compared to the control. Consequently, only the European trial suggested that there is a benefit in prostate cancer mortality due to early detection and treatment.

Such divergences in the two studies results contributed to that currently there is not a global agreement in effectiveness of prostate cancer screening. Moreover, recent studies (Chou et al., 2011; Carter et al., 2013) are increasingly focusing on the possible adverse effects of treatment for prostate cancer, which were reported in none of the mentioned studies. These adverse effects need to be balanced against any potential reductions in mortality.

Hence, organizations such as the *European Association of Urology* (EAU), or the *United States Preventive Services Task Force* (USPSTF), have advised against introducing a mass screening programme for prostate cancer, after a review revealed that the harms involved with using the PSA test to screen for the disease would outweigh the benefits. This decision was motivated by the fact that *PSA* test by itself is not a highly sensitive test, so that we are unable to correctly identify those cancers which will progress and those which are indolent and may be safely watched without treatment.

All these facts explain why the governments do not apply population programs for early detection of the prostate cancer. However, although the benefits of screening are not at all clear, recent trends indicate that screening may be a valid methodology provided that there exists a physician–patient relationship that allows to understand the harms and potential benefits. In this line, (Moyer, 2012) admitted the existence of a very small potential benefit, so it encouraged clinicians to screen only the men who previously had been informed about the known risk of harms. This recommendation is framed within the current trend towards more personalized medicine, taking the peculiarities of each patient into account.

2.2 European Randomized Screening for Prostate Cancer: ERSPC study

2.2.1 Motivations and aims of the ERSPC

In recent decades, there have been many medical studies focused on the development and implementation of statistical techniques aimed at identifying the most critical covariates in the diagnosis of a particular disease, with great emphasis on oncological diseases. Thereby, the interest in predicting the risk of cancer is part of the current personalization trend of medical interventions, which in this case focuses on early disease detection through a screening process.

As already mentioned in the previous section, we can found some evidences showing that screening for certain cancers can reduce mortality and improve patients' quality of life. Screening can also cause side effects in the healthy population in terms of false positive or false negative results, as well as problems related to overdiagnosis. The problem basically consists of diagnosting cancers (with all that implies in treatment terms) in individuals who would otherwise have died of natural causes without a clinical diagnosis of cancer.

Thus, it is essential to reduce the enormous costs, both personal and financial, resulting from incorrect treatment of the subject. To do this, it is necessary a further deepening of the statistical treatment of such data, developing and implementing new models to improve cancer prediction before symptoms. In this direction, the most scientifically valid way to assess the effect of early detection of disease is still by means of a randomized controlled trial with death caused by the disease as the main endpoint.

Among the contributions made in improving prediction of oncologic diseases, prostate cancer is a disease whose early diagnosis may contribute to reduce the mortality rate. In this specific task, in the early 1990's began the aforementioned ERSPC project. It is a large randomized trial of screening for prostate cancer, consisting of comparing an intervention arm (men to whom regular PSA screening is offered), with a control arm (men to whom such screening process is not done).

The general goal of the ERSPC trial was to evaluate whether prostate cancer screening reduces mortality from this disease in the asymptomatic population, within the scope of the different countries of the *European Random Screening for Prostate Cancer* study, (ERSPC). Consequently, **ERSPC represented a survival study in which the event of interest was death caused by prostate cancer** (Schröder et al., 2012).

The particular aims of this study were the following:

1. Assess the extent to which false-positive screening results occurred in repeated prostate cancer screening in the whole ERSPC trial.

2. Determine the possible risk for future prostate cancer, another false-positive screening result and subsequent non-compliance in screening in men with false-positive screening results.

3. Establish how screening affects the prostate cancer incidence (especially advanced cancer) in Europe.

4. Determine how screening affects prostate cancer mortality in Europe, and which factors contribute most to the group of screening failures (i.e. men who die of prostate cancer in spite of screening).

2.2.2 Recruitment, randomization and protocol in the ERSPC

Pilot studies in Belgium and the Netherlands (the first participating centers of ERSPC) were conducted between 1991 and 1994 and the results were reported. The conduct of similar pilot studies became a condition for the admission of other European centers that joined the ERSPC in subsequent years. Actually, the ERSPC study consists of seven countries having a sufficiently long time monitoring: Belgium, Finland, Italy, Netherlands, Spain, Sweden and Switzerland (France joined up to study in 2003).

Recruitment and randomization procedures differed among the study countries, having been developed in accordance with national regulations. On one hand, in Finland, Sweden, and Italy, the trial subjects were identified from population registries and underwent randomization before written informed consent was provided (population-based effectiveness trial). On the other hand, the target population of Belgium, the Netherlands, Spain and Switzerland was also identified from population lists, but when the men were invited to participate in the trial, only those who provided consent underwent randomization (efficacy trial).

The seven original participating countries in ERSPC trial and the specific characteristics of each study section are illustrated in Table 2.1.

Country	Recruitment	Age at Entry (yr)	Screening Began	Interval (yr)	*PSA* cut-off (ng/ml)
Belgium	Volunteer	50−74	1991−2003	4−7	*PSA* ≥10 (1992−1994) *PSA* ≥4 (1995−1997) *PSA* ≥3 (from 1997)
Finland	Population	55−67	1996−1999	4	*PSA* ≥4
Italy	Population	55−71	1996−2000	4	*PSA* ≥4
Netherlands	Volunteer	55−75	1991−1999	4	*PSA* ≥4 (1991−1997) *PSA* ≥3 (from 1997)
Spain	Volunteer	45−71	1996−1999	4	*PSA* ≥4 (1996−1998) *PSA* ≥3 (from 1998)
Sweden	Population	50−66	1994	2	*PSA* ≥4
Switzerland	Volunteer	55−70	1998−2003	4	*PSA* ≥3

* French data are not included

Table 2.1. Characteristics of the screening protocols of the seven older ERSPC members (Adapted from ERSPC Conclusions, Schröder et al. (2012).

In the screening arm, tests were performed every 4 years with the exceptions of Sweden, where rescreening took place after 2 years, and Finland, with exceptional rounds of 7 years during certain period. In all cases, therapy started when cancer was detected. In contrast, in the control arm no tests were performed.

At the beginning of 2004, the study had recruited a total of 182 160 asymptomatic men, aged

45−75 years and distributed among the following seven countries: Belgium, Finland, Italy, The Netherlands, Spain, Sweden and Switzerland. With all these countries taking part in this project, it became the world's largest prostate cancer screening.

From the total number of recruited men among the seven chosen countries, 160 died during randomization process, finally remaining a total of 182 000 men. Their distribution among countries is displayed in Table 2.2.

Country	Men Recruited	Randomization	
		Screening Arm	Control Arm
Belgium	9980	5000	4980
Finland	80458	32000	42458
Italy	14557	7286	7271
Netherlands	42376	21210	21166
Spain	4278	2416	1862
Sweden	19946	9973	9973
Switzerland	10405	4931	5475
Total	**182000**	**82816**	**99184**

* French data are not included

Table 2.2. Recruited men in the first seven ERSPC members (Adapted from ERSPC Conclusions, Schröder et al. (2012)).

These recruited men were divided between screening and control groups as showed in Figure 2.2:

a) 82 816 men were assigned to screening arm, of whom 82% had at least one PSA test during trial.

b) 99 184 men were assigned to the control group. Based on single site, screening in controls estimated at approximately 20%.

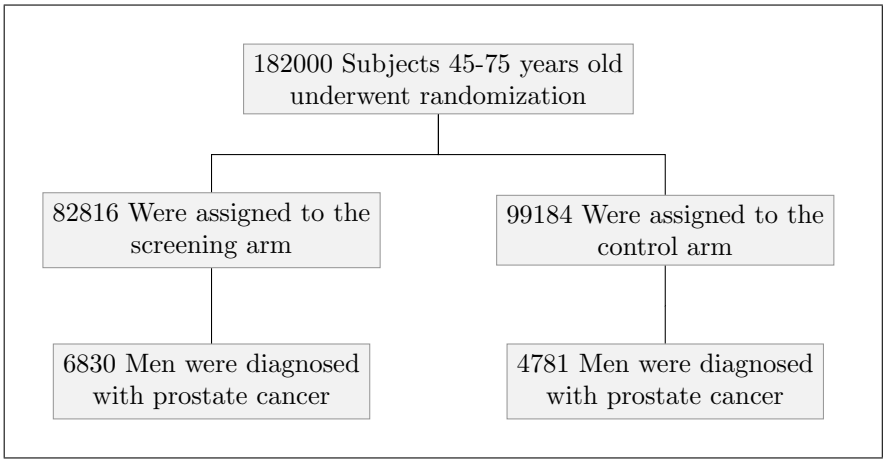

Figure 2.2. Enrollment and outcomes according to age group at randomization (Adapted from Schröder et al. (2012))

2.2.3 General conclusions of the ERSPC

In both arms, similar prostate cancer cases were treated similarly. After a mean follow-up of more than 10 years, in the screening group 6830 individuals were diagnosed with prostate cancer, whereas

in the control group were diagnosed 4781 cases (Schröder et al., 2012). These data are unlikely to help to clarify the debate over the value of mass screening for prostate cancer. From one point of view, one could use them to argue that screening can prevent between 20 and 30 percent of prostate cancer-specific deaths with only 9 years of follow-up. From the alternative point of view, one could argue that screening a million men would indeed prevent about 950 prostate cancer-specific deaths, but would also lead to the potential over-treatment of 36 out of every 37 cases of prostate cancer identified, and would have no impact whatsoever on overall mortality (Roobol et al., 2012).

In general, prostate cancer screening has remained controversial because it has also led to considerable false positive results and extensive overdiagnosis of disease that would not otherwise emerge clinically. In the case of the two large studies cited, ERSPC trial revealed a significant reduction in the rate of death from prostate cancer (relative reduction, 20%) among men offered screening for prostate-specific antigen. In contrast, the American PLCO, did not show an effect of screening on prostate cancer mortality. Furthermore, the critics of screening programs argue that in ERSPC project, the benefit of screening is associated with an unacceptably large proportion of overdiagnosis and potential overtreatment. This divergence of results leads to the fact that screening utility is still under scientific discussion, indicating that shared decision making is necessary prior to testing and taking prostate biopsies.

2.3 The Spanish section of the ERSPC

2.3.1 Development of the ERSPC Spanish section

The study population in the ERSPC Spanish branch (Berenguer et al., 2003; Luján et al., 2012) comprises men recruited from two cities, Getafe and Parla, located at the *10th Health Area* from Comunidad de Madrid. After a mailed invitation to 18612 men of the referred health zone, the Spanish section finally enrolled 4278 individuals between February 1996 and June 1999 (participation rate of 23%). At the time of randomization, these men were 45 to 71 years old and had a life expectancy greater than 10 years. Although randomization was then made 1:1 to either screening (tests with *PSA*) or control group (no tests), the number of subjects recruited was different in both arms due to some initial quirks in the recruitment process (by decision of the *Quality Control* group, there were suppressed those control arm subjects who were not finally invited). For this reason, among the 4278 men who were recruited, 2416 subjects were assigned to the screening arm and the remaining 1862 to the control arm.

Year	Men Invited	Men Accepted	Randomization	
			Screening Arm	Control Arm
1996	1190	268 (23%)	124	144
1997	4036	1124 (28%)	850	274
1998	11372	2684 (24%)	1344	1340
1999	2014	202 (10%)	98	104
Total	**18612**	**4278 (23%)**	**2416**	**1862**

Table 2.3. Recruited population by year in ERSPC Spanish section (Berenguer et al., 2003).

The database of the ERSPC Spanish section included several variables like subject identifier, date of birth, date of randomization, arm of study (screening or control), dates of attendance (*PSA* testing), biopsy result or prostate cancer detection. In both arms, annual mortality was studied

and cause of death recorded thanks to the agreement reached with the *Spanish Institute of Statistics* (providing the date of death and its underlying cause, as stated in the death certificate).

During the study it was performed a follow-up of the screening arm subjects, recommending a biopsy based on the only criterion of the serum *PSA* level registered in a visit of the individual. In particular, biopsies were indicated when *PSA* level was above a pre-specified threshold, and if the result was negative then men were invited to a new re-screening round after a year interval ("early recall"). However, if the *PSA* level was considered normal, the standard interval to the next screening round ("routine recall") was 4 years. In contrast, no tests were applied in men allocated to the control group.

Subjects in the screening group have a variable number of visits. In each visit the patient's total *PSA* level evolution was recorded, and from 2002 onwards, it was also registered the free *PSA* (*FPSA*, ng/ml) and *PSA* ratio (defined as $PSARAT = FPSA/PSA$). Furthermore, on some dates distributed throughout the study patients also had a digital rectal examination (*DRE*) and an ultrasound rectal test (*TRUS*). However, no biopsy was indicated based on *DRE* or *TRUS* findings.

In the screening group, *PSA* thresholds for which the study recommended the biopsy varied over time, distinguishing three periods:

1. First protocol: From February 1996 to April 1998:

   Biopsy prescription for those individuals with a $PSA > 4$ ng/ml, and in the case of negative result the patients were called to an early recall visit on a time period under two years. Moreover, the protocol also established that if *PSA* level recorded during a visit was such that 3 ng/ml $\leq PSA \leq 4$ ng/ml, then a biopsy test was not prescribed but the patient was called for a new visit within 2 years.

2. Second protocol: From May 1998 to December 2001

   Biopsy indicated to anyone who registered $PSA \geq 3$ ng/ml. In the case of a negative result on biopsy, the person would be called to an early recall (always within less than two years), undergoing further evaluation ("rescreen") that included a new determination of *PSA* (and indicating again a biopsy if *PSA* level exceeded the allowable threshold).

3. Third protocol: From January 2002 to October 2005

   The protocol used the same criterion as in the previous period plus a biopsy indication when the total *PSA* was in the range $1-2.99$ ng/mL and $PSARAT \leq 0.20$.

First visits with *PSA* readings took place on 19th February 1996, and from that date began a three-year period of first visits that lasted until 30th June 1999. Hence, the screening group men must have his first visit located within this three-year period. Last visits with *PSA* measurements occurred in October 2005, and the database includes active surveillance of prostate cancer incidence until 31/12/2007.

Accordingly with the above information, **the study started at 19th February 1996, and the finalization of active surveillance on prostate cancer incidence took place on 31st December 2007.**

2.3.2 Specific results for the ERSPC Spanish section

There were recruited 2416 subjects in the screening arm and 1862 in the control. As indicated by Luján et al. (2012), their average age was 57.8 years and the median follow-up period 13.3 years. At

the end of follow-up were recorded a total of 427 deaths (the event of interest in ERSPC project), of which 9 were related to prostate cancer.

Regarding the causes of death, the main cause was due to malignant tumors (52.9%), highlighting the lung (15.9%) and colorectal cancer (7.0%). Secondly, cardiovascular diseases(17.3%) and respiratory incidences (8.9%). In addition, no differences were observed in the distribution of the causes of death between the control and screening (p-value = 0.20).

Survival analysis (consisting of non parametric estimation) was used to compare mortality rates between screening and control groups during the 15-year period of the study, and a log-rank test was performed between survival estimated curves. The results indicated no significant differences between the two arms with a highly evidence that led to not rejecting the null hypothesis (p-value = 0.94). In the same line, there was no significant effect between arms when performing a comparison based on mortality rates due to different cancer types (p-value = 0.54).

Thus, the study results indicated that we could not establish differences in mortality after 15 years of follow-up. However, the prostate cancer mortality in the population at the end of this period was very limited (less than 1%).

2.4 Source dataset of the cohort under study

2.4.1 Source dataset: The Prostate Cancer (PCa) Dataset

The aim of this section is to introduce the dataset that has motivated the different methodologies applied in this research work. In particular, information derived from ERSPC Spanish section has been used, keeping in mind all variables which are expected to have effect during the different analyses. The database was called the **PCa Dataset**.

Our source dataset has been configured to give an exhaustive picture of all data provided by the ERSPC Spanish branch, despite that the `PCa Dataset` compiles information which has not been directly used in the subsequent chapters of this work research, either because a lack of longitudinal follow-up or because the need of implementing other more advanced methodologies.

Our study will focus specially in the subject's *PSA* level at their respective time points, because this biomarker is present in each of the records and is the main parameter to recommend if the patient should undergo a biopsy test. As already mentioned before, a biopsy test is ultimately the gold standard to establish whether there is a diagnosis of prostate cancer.

2.4.2 The double target in PCa Dataset configuration

The `PCa Dataset` allows to model the given information from a simultaneous dual approach:

- Survival approach: The `PCa Dataset` contains the follow-up time from the first visit of each subject until the possible detection of prostate cancer. This information allows to determine the observed survival time from the birth date of each subject up to the mentioned event, relating the event time outcome to the time-dependent covariates in the studio. In contrast with the ERSPC study, where the time until subject's death by prostate cancer was analyzed, **in our study the event of interest, $\mathcal{E}$, is defined as the diagnosis of prostate cancer**.

- Longitudinal approach: Secondly, it must be taken into account that both the true importance of a biomarker in describing prostate cancer progression and its association with survival time can

only be revealed when repeated measurements of the marker are considered in the analysis. For this reason, the **PCa Dataset was also constructed in order to carry out a longitudinal study of different time-dependent covariates associated with each subject**, allowing us to establish the temporal relationship between them and with respect to the particular event of interest.

Both commented approaches are different but complementary. The ultimate goal is to obtain a dataset in a way that allows us merging both survival and longitudinal processes into a one statistical model, giving rise to the so-called *joint models*, JM. Such models allow to establish the association between a longitudinal biomarker and time until some specific event occurs.

2.4.3 Information provided by the PCa Dataset

The `PCa Dataset` included the 2416 subjects assigned to the screening arm of the Spanish section ERSPC study. One individual from the screening arm was removed from the dataset for having had a previous cancer, so **we finally had information in our `PCa Dataset` on 2415 men, representing 4673 visits which are unequally distributed among them**.

For each subject, the individual profile description covered from his first visit date until the date on which the first of these three possible scenarios happens:

- Scenario 1: The subject profile reached the study end, on 31/12/2007, without being diagnosed with prostate cancer, that is, without having developed the disease.

- Scenario 2: The subject profile ended at the date of his death, prior to 31/12/2007. Although in this case the individual did not achieve the end of the study, he did not experience the disease during the study.

- Scenario 3: The subject profile ended at the date of his prostate cancer diagnosis, prior to 31/12/2007. In this case the individual developed the disease within the study.

In considering the three scenarios described above, it is clear that the treatment of the data must be different depending on the case. In particular, in either of the first two scenarios the individual did not experience the event of interest, while in the the third scenario the profile of the individual did reach this event. Consequently, in the first two cases we could only say that up to a determined time point, the subjects *survived* the disease occurrence, but we had not more information beyond that date.

The statistical inference with survival data is usually complicated by presence of these incomplete observations. This distinction has its origin in the phenomenon known as right-censoring data, consisting of some individuals not experiencing the event of interest when they reached the closing date of the study or dropping out of the study before reaching that date. The analytic treatment of right-censored data and its implications will be discussed in detail in the next chapter.

Once the general casuistry was defined, the interest lied in able to define a dataset that allowed not only good survival characterization, but also a longitudinal one. Among many other variables that were collected in the Spanish Section of ERSPC study, the following 16 variables were included in the `PCa Dataset`:

Name	Description
OBS	Row identification
ID	Subject identifier
RANDAT	Date of randomization
DATE	Date of record
RECTYP	Type of record, categorized for each observation according to the code: 0: Observation associated with a record date that refers neither to a trial exclusion of the subject nor at the time of his death. 1: Observation associated with a record date of exclusion from the study of the subject. The "exclusion" means that the subject has not been invited to further re-screening, but the study has continued to gather information on his potential prostate cancer incidence until 31/12/2007. 2: Observation associated with the death date of the patient, which means that the subject has died before 31/12/2007.
AGE	Age of the patient at each of his record dates (years). The age was obtained in days and divided by 365.25
PSA	Total level of prostate-specific antigen (ng/ml)
DRE	Information available on digital rectal examination 0 = Not performed, 1 = Negative, 2 = Positive, NA = Unknown
TRUS	Information available on transrectal ultrasound test 0 = Not performed, 1 = Negative, 2 = Positive, NA = Unknown
TIME	Observation time (years), which may correspond to an event or to a right-censored data.
CENS	Censoring indicator, categorized for each subject according to the code: 0 = Right-censored, 1 = Event
BIOPSY	Biopsy result, categorized for each observation according to the code: 0 = Not recommended, 1 = Not performed but recommended, 2 = Negative result, 3 = Positive result: Screening prostate cancer diagnosed, NA = Unknown
FPSA	Free *PSA* level (ng/ml). Result = real positive number (0.nn) , NA = Unknown
PSARAT	*PSA* ratio, defined by *PSARAT* = *FPSA/PSA* (range between 0 and 1). Result = real positive number (0.nn), NA = Unknown
CANTYP	Type of cancer, categorized for each subject according to the code: 0 = Not tumor, 1 = Screen-detected cancer, 2 = Interval cancer, NA = Unknown
GLE2	Gleason grading system, which is measured in a 1−3 scale, obtained for each subject by rescaling (according to the ERSPC criteria) the Gleason results (1−10). If Gleason = 0 then *GLE2* = 0: Not performed (no tumor) If Gleason in 1−4 then *GLE2* = 1: Low aggressiveness level If Gleason in 5−7 then *GLE2* = 2: Medium aggressiveness level If Gleason in 8−10 then *GLE2* = 3: High aggressiveness level If Gleason = NA then *GLE2* = NA: Missing value (unknown)

Table 2.4. Names and description of the variables in the `PCa Dataset`.

Due to the time depending covariates on a given subject, the data was arranged in the so-called *long format*, in which the measurements of each individual are stored in multiple lines. Thus, the `PCa Dataset` contains in this long format both the survival and the longitudinal information of

each subject.

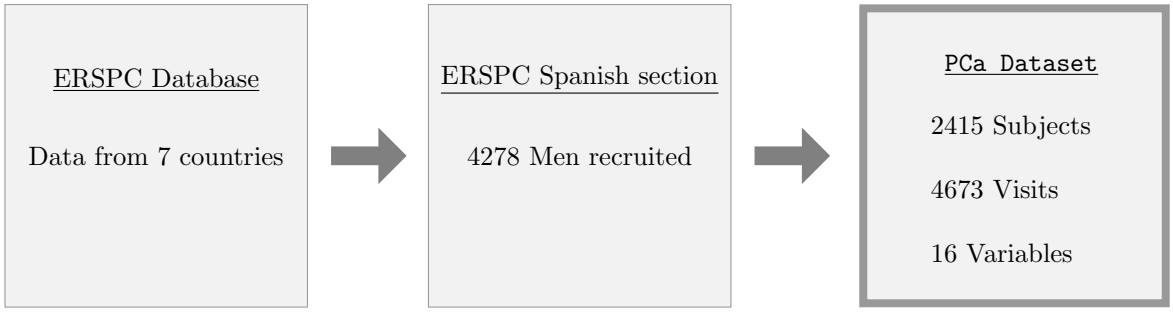

Figure 2.3. Obtainment of the `PCa Dataset` from the ERSPC Spanish branch data.

Table 2.5 displays the format and the structure of the `PCa Dataset`, and more information can be found in Appendix 1.

ID	*RANDAT*	*DATE*	*RECTYP*	*AGE*	*PSA*	*DRE*	*TRUS*	*BIOPSY*	*FPSA*	*PSARAT*	*TIME*	*CENS*	*CANTYP*	*GLE2*
1	16/04/1996	26/04/1996	0	54.09	0.70	0	0	0	NA	NA	65.77	0	0	0
1	16/04/1996	25/10/2000	0	58.59	0.60	0	0	0	NA	NA	65.77	0	0	0
1	16/04/1996	25/02/2004	0	61.92	0.68	0	0	0	NA	NA	65.77	0	0	0
9	03/01/1998	03/02/1998	0	54.86	0.70	0	0	0	NA	NA	64.77	0	0	0
10	05/04/1996	15/04/1996	0	53.06	0.20	0	0	0	NA	NA	64.77	0	0	0
10	05/04/1996	03/10/2000	0	57.53	0.10	0	0	0	NA	NA	64.77	0	0	0
10	05/04/1996	24/02/2004	0	60.92	0.19	0	0	0	NA	NA	64.77	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9478	23/06/1999	23/06/1999	0	55.30	0.00	0	0	0	NA	NA	63.82	0	0	0
9478	23/06/1999	24/10/2003	0	59.64	0.69	0	0	0	NA	NA	63.82	0	0	0

Table 2.5. Layout information in `PCa Dataset`.

2.4.4 Description of the PCa Dataset

Study duration

The study began on 19th February 1996 (visit date of the first men to join the study) and ended, due to administrative reasons, on 31st December 2007. The period of time between two dates is called *observation window* or sometimes *time window*, and corresponds to the time period when it was performed an active surveillance of the 2415 dataset subjects.

Longitudinal characterization

The `PCa Dataset` contains 2415 individuals aged 45−71 years from the screening arm of the ERSPC Spanish section. In Figure 2.4 is displayed how the mean age at randomization process, conducted during the first three years of the study, was 57.72 years and the standard error 5.45 years (Luján et al., 2012).

The range of follow-up in time was 0.00−11.86 years, with a mean of 9.28 years and a median of 9.73 years. At this point, it must be noted the limited longitudinal monitoring of the data, as evidenced by the fact that the average number of records by subject is about two measurements (Table 2.6).

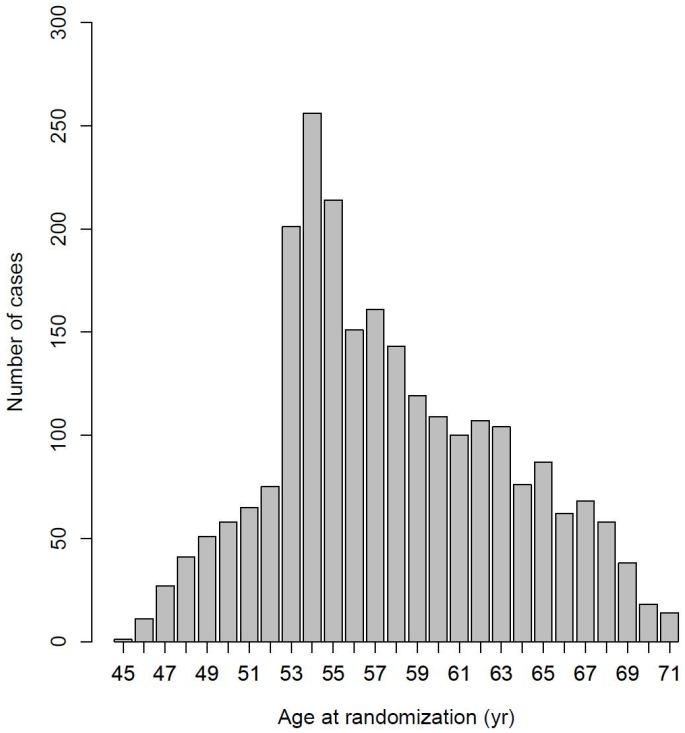

Figure 2.4. Age of 2415 subjects at their randomization date in the ERSPC Spanish section (Luján et al., 2012).

This fact had imposed a logic limitation in getting a more precise characterization of individual trajectories, and must be taken into account when interpreting the study results. However, the small number of time points for each individual does not mean that the obtained results obtained are wrong: simply the results accuracy would increase with a greater number of records by subject.

Distribution of the number of visits	
Subjects with 1 visit	573 (23.7%)
Subjects with 2 visits	1499 (62.1%)
Subjects with 3 visits	293 (12.1%)
Subjects with ≥ 4 visits	50 (2.1%)
Average number of visits per subject	1.9

Table 2.6. Distribution of the number of visits among the 2415 subjects in the `PCa Dataset`.

Information about explanatory covariates

As previous commented, visit study's principal covariate is the *PSA* level in each visit. Serum *PSA* determinations were mostly performed at an interval of 4 years. A biopsy was indicated if $PSA \geq 3$ ng/ml, and in case of a negative result the subject was recalled for an early visit within a 2 years period. Globally, among the 2415 participants, a total of 4673 follow-up *PSA* measurements at the successive screening visits have been recorded. Last visits with *PSA* readings in the study took place in October 2005, and the survey data include active surveillance of prostate cancer incidence until 31st December 2007, not taking into account any collected information about the subjects occurred after that date.

Concerning the number of visits per subject, data are not balanced because individuals do not enter the study at the same age. Table 2.7 shows the distribution of subjects according with the number of visits and a descriptive statistics of the subjects' *PSA* mean measurements for each one of the categories and the overall sample, as well as the frequency and percentage of prostate cancer diagnosed subjects. In this regard, we can remark the following important results: a) the subset of men with two *PSA* measurements is the most frequent (62.1%), b) there is 23.7% of subjects that they are contributing to the analysis with only one *PSA* observation, scarcely nourishing the longitudinal analysis, c)this group with only one visit contains individuals with a huge variability of *PSA* level (StDev = 5.67 ng/ml and Max = 69.80 ng/ml), and d) categories with one or two visits concentrate the main proportion of PCa diagnosed cases (87.1%).

	Number of visits per subject				
	1	2	3	≥ 4	Overall
Sample size (%)	573 (23.7)	1499 (62.1)	293 (12.1)	50 (2.1)	2415 (100.0)
PSA descriptive					
Mean	2.40	1.40	1.92	5.46	1.78
StDev	5.67	1.67	2.25	2.79	3.20
Min	0.00	0.00	0.00	0.50	0.00
1st Q	0.60	0.60	0.60	3.79	0.63
Median	1.10	1.00	1.10	4.69	1.06
3rd Q	2.20	1.67	2.44	6.28	1.95
Max	68.90	42.30	20.36	15.53	68.90
PCa diagnosed (%)	51 (44.0)	50 (43.1)	13 (11.2)	2 (1.7)	116 (100.0)

Table 2.7. Distribution of the subjects, descriptive statistics of the *PSA* measurements and distribution of the prostate cancer diagnosed cases, stratified by the number of visits and overall.

On the other hand, the results for categorical covariates *DRE* and *TRUS* were also collected at each individual's visits, which in the `PCa Dataset` play the role of auxiliary variables that help to calibrate *PSA* effect on prostate cancer risk.

Table 2.8 presents a descriptive summary of *PSA* levels stratified by combined *DRE* and *TRUS* categories:

DRE	*TRUS*	Visits	Subjects	Events	Subj./Ev.(%)	*PSA* Summary						
						Mean	sd	Min	1stQ	Med	3rdQ	Max
0	0	3999	2300	14	0.61	1.26	1.46	0.00	0.59	0.90	1.50	47.00
0	1	16	16	1	6.25	3.96	1.37	1.19	3.18	3.67	4.20	7.12
0	2	1	1	0	0.00	14.00	–	14.00	14.00	14.00	14.00	14.00
1	0	45	45	6	13.33	4.13	3.51	1.05	1.70	3.50	5.00	19.30
1	1	500	382	60	15.71	4.55	3.41	1.00	3.00	4.07	5.40	51.60
1	2	33	32	14	43.75	7.67	4.50	1.60	4.40	6.00	9.40	21.70
2	0	12	12	2	16.67	2.86	1.42	1.25	1.80	2.77	3.65	6.10
2	1	40	37	8	21.62	5.70	3.51	1.19	3.20	4.33	7.20	13.20
2	2	27	26	11	42.31	14.86	18.86	1.71	3.90	5.60	13.90	68.90

Table 2.8. Distribution of the subjects, descriptive statistics of *PSA* measurements and distribution of the PCa diagnosed cases, stratified by *DRE* and *TRUS* values at each of the 4673 visit dates.

The above table displays that most observations correspond to the pair ($DRE = 0$; $TRUS = 0$), so not performed tests represent a total of 85.6% over 4673 recorded time points. It is also

interesting to note that 2300 subjects had at least one *PSA* level measurement associated with this pair. Considering the informative character of the two auxiliary covariates (the zero value can be associated to a low risk *PSA* level), it can be concluded that most of the studied subjects will move around low *PSA* values, having therefore a low risk of developing prostate cancer.

High right-censored data proportion

At the end of follow-up 116 (4.8%) prostate cancers were diagnosed, while the remaining 2299 (95.2%) subjects led to right-censored data. This result is in agreement with the observations made previously, showing that it is a disease with a low incidence rate over the total adult population.

However, the fact that the incidence is low does not imply that it is irrelevant. A percentage of 5% in cancer incidence during the window observation is high enough to be considered a detailed study to identify what are those more important risk factors in the disease development. In this regard, it would be desirable to have information about another important risk factors like family history, prostate volume, etc.

2.4.5 Transformation of some variables from PCa Dataset

Double log-transformation of *PSA* variable: *LLPSA*

In the health research studies, many of the most classical biomarkers are continuous variables that have heavily skewed distributions, with long right tails. We can have one or a few values which are much larger than the vast majority of the data, so that these high values make frequently very difficult to identify the structure in the rest of the data. Furthermore, these kind of variables are not normally distributed, and therefore do not satisfy the required assumptions to apply neither certain linear models nor parametric statistical tests.

In connection with the above paragraph, the *PSA* biomarker is a clear example of biological variable with high skewed distribution. In particular, its values stand typically under 5 ng/ml, but can be placed above 50 ng/ml (outliers at the high end). In our study, the *PSA* ranged from 0.00 to 68.90 ng/ml with mean and median values 1.78 and 1.06 ng/ml., respectively (Table 2.7). There is therefore a need for a mathematical transformation of *PSA* values to avoid masking the structure presented by smallest values.

Considering the longitudinal character of the *PSA* variable, the change experienced in applying successive logarithmic transformations should be undertaken at a given time point. Moreover, it must be taken into account that not every men have the same number of visits (and therefore of measurements), and each of them contributes with different time points. It was consequently decided to use the *PSA* value at their respective first visits as a common reference point, then assimilating that measurement to a basal value denoted by PSA_0.

To accommodate for data normality, it was originally considered a non linear transformation based on application of a logarithmic scale. Therefore, larger values moved closer together while smaller values stretched out, and the original values reduced to a more manageable size. We also had to take into account the presence of null values, so we decided to add a constant $k = 1$ as proposed in Slate and Turnbull (2000), so in our case: $LPSA_0 = \log(1 + PSA_0)$. However, the dispersion of our *PSA* measurement values was so high that a double logarithmic transformed scale was needed: $LLPSA_0 = \log\{1 + \log(1 + PSA_0)\}$.

As showed in Figure 2.5, the successive log-transformations of *PSA* variable make positively skewed

distribution more normal. In the original scale, the data are long-tailed to the right, but after a double logarithmic transformation is applied, the data distribution tends to be symmetric.

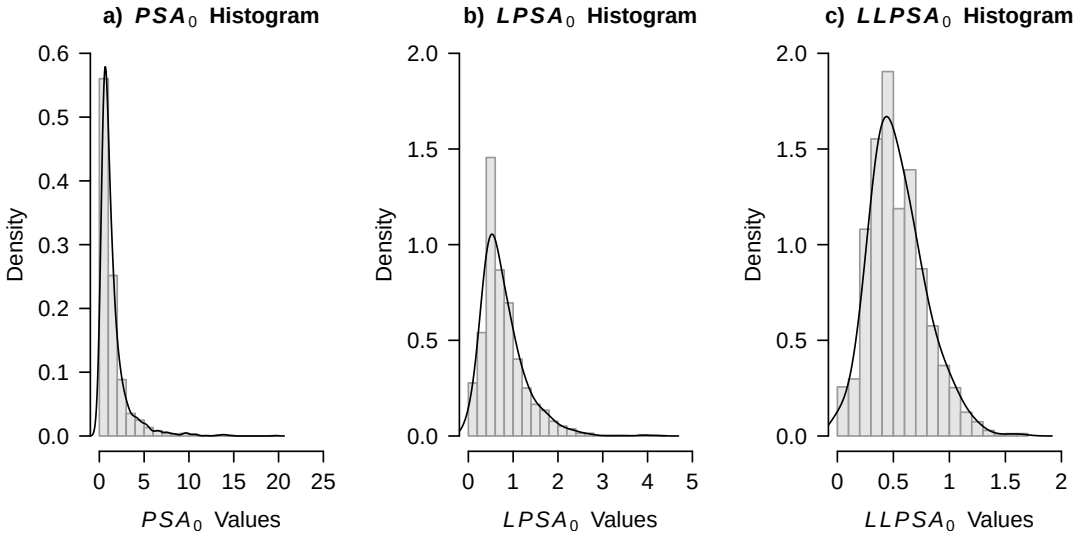

Figure 2.5. (a) Original PSA_0 data histogram, (b) PSA_0 data histogram after first log-transformation: $LPSA_0$ data, and (c) PSA_0 data histogram after double log-transformation: $LLPSA_0$ data.

Hence, **the main variable used in this prostate cancer study has been *LLPSA*, defined as:**

$$LLPSA_{ij} = \log\{1 + \log(1 + PSA_{ij})\} \tag{2.1}$$

for the $i-$th subject, $i = 1, \ldots, 2415$, at t_{ij} time point, $j = 1, \ldots, n_i$

Translation of the starting point

It is important to note that all men in the study were at least 45 years old at the time of first visit. In order to avoid working with a long period of time without data, we opted to apply a 45 years-translation, and considered 45 years as the initial time point:

$$\begin{aligned} AGE'_{ij} &= AGE_{ij} - 45 \\ TIME'_{ij} &= TIME_{ij} - 45 \end{aligned} \tag{2.2}$$

for the $i-$th subject, $i = 1, \ldots, 2415$, at his t_{ij} time point, $j = 1, \ldots, n_i$

Relation between the *LLPSA* and *AGE* covariates at first visit

Since time from study entry to prostate cancer diagnosis can vary depending on the age, it is important to account for this issue.

Figure 2.6 displays $LLPSA_0$ histrograms, stratified by tertiles of age at entry ([45; 55), [55; 60) and ≥ 60 years), namely AGE_0. We can realize that, when age increases, the $LLPSA_0$ levels move toward higher values.

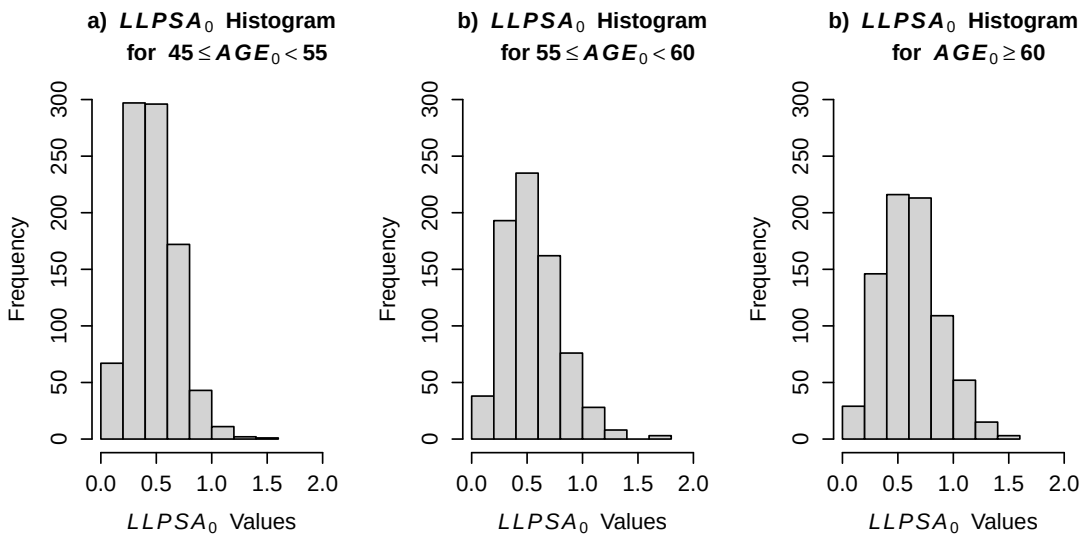

Figure 2.6. $LLPSA_0$ data histogram for (a) $45 \leq AGE_0 < 55$, (b) $55 \leq AGE_0 < 60$, and (c) $AGE_0 \geq 60$.

The existing interaction between both variables results in the need to analyze the normality of $LLPSA$ on the basis of its dependence on AGE. For this purpose, the first visit values were considered again as a baseline, so we worked in terms of $LLPSA_0 \times AGE_0$. The assumption of a normal model for a population of responses will be required in order to perform certain inference procedures, and histograms could be used to get an idea of the shape of a distribution. However, there are more sensitive tools, like the quantile-quantile plot test, commonly called *q-q plot*. In general, the q-q plot goal is to verify the assumption of normality, comparing the distribution of the sample to a normal distribution. The plot represents the quantiles of a standard normal distribution against the corresponding quantiles of the observed data. If the observations follow approximately a normal distribution, the resulting plot should be roughly a straight line with a positive slope (consequently, deviations from this line would indicate possible departures from a normal distribution).

Figure 2.7 presents the q-q plots associated to each 9 possible combinations according to the degree of transformation of PSA_0 values and stratifying by AGE_0. This multi-panel q-q plot displays an underlying normality as logarithmic are applied to PSA_0 values. Thus, the q-q plots corresponding to $LLPSA_0$ show evidence of an approximately normal distribution except for some large outliers which in any case do not change the general behavior.

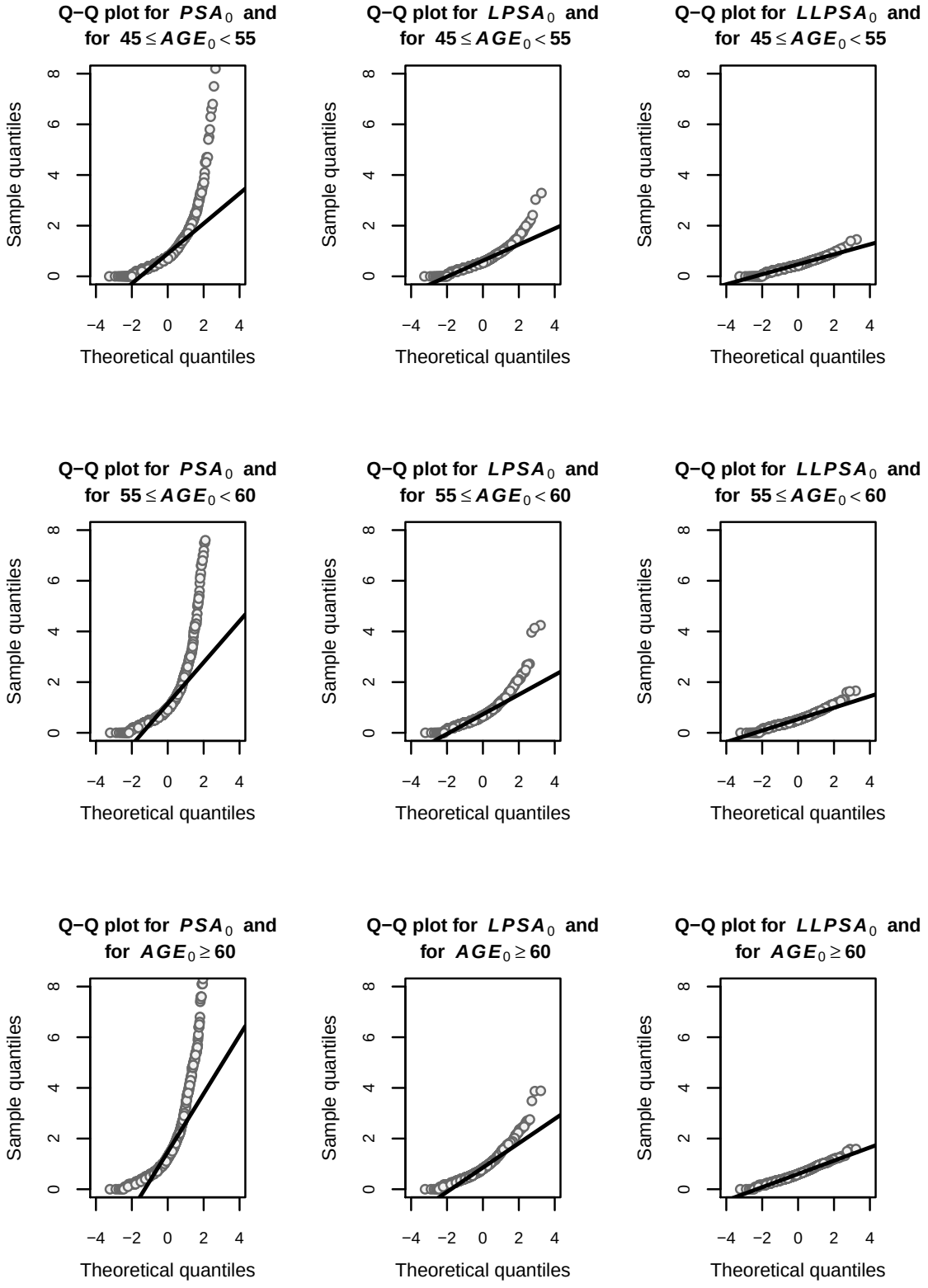

Figure 2.7. Normal quantile plots depending both on the PSA_0 subject's level and his respective AGE_0 tertile. The three left panels correspond to q-q plots of the untransformed PSA_0 data, the middle column panels contain the q-q plots for log-transformed data, and the three right panels are the q-q plots for double log-transformed measurements. From the top row to the bottom row, the AGE_0 tertile increases.

CHAPTER 3

JOINT MODELING OF LONGITUDINAL AND SURVIVAL DATA

3.1 Motivations for Joint Modeling

Many current investigations generate both longitudinal measurement data, with repeated measurements of a response variable at a specific number of time points, and time-to-event data, in which times until a particular event are recorded. For example, in many medical studies, we often collect patients' information repeatedly over time, and we are also interested in the time to recovery or recurrence of a disease. Separate analyses of longitudinal and survival data may lead to inefficient or biased results, so it is necessary a jointly modeling in order to incorporate all information simultaneously and provide valid and efficient inferences. Such an approach is termed *Joint Modeling*.

Joint models of longitudinal and survival data have attracted increasing attention over the last two decades. They were introduced during the 90's (Tsiatis et al., 1995; Faucett and Thomas, 1996; Wulfsohn and Tsiatis, 1997) and since then have been applied to a great variety of studies in epidemiological and biomedical areas. In turn, these studies have fed a wide methodological research on the subject, with models focused on event times, longitudinal patterns or both. In this sense, Tsiatis and Davidian (2004) provide excellent review up to date. More recently, Rizopoulos has made a great contribution facilitating the use of the joint modeling methodology, first by means of an excellent overview of the theory and applications of joint modeling (Rizopoulos, 2012b) and secondly by developing the `JM` (Rizopoulos, 2010) and `JMbayes` (Rizopoulos, 2012a) `R` packages for the frequentist and Bayesian approaches, respectively. In particular, computational applications included in this work is based on Rizopoulos' contributions.

A typical joint model setting is to assume a linear mixed effects model for the longitudinal covariates and a Cox model or an accelerated failure time (AFT) model for the survival data, with the two models sharing some random effects or covariates. Therefore, by joint modeling techniques longitudinal measurements are adjusted to allow for non-ignorable missing data due to informative dropout, which cannot be appropriately handled by the standard linear mixed effects models alone. For this purpose, the likelihood method is often used, implemented by EM algorithms (Dempster et al., 1977).

3.2 Longitudinal Data Analysis

3.2.1 General features of longitudinal data

Longitudinal studies are based on the data resulting from the measurements of subjects taken repeteadly at multiple follow-up times, thereby allowing the direct study of changes in the response variable within the duration of the study. Such data are frequently encountered in health sciences, in which longitudinal studies play a prominent role in enhancing the understanding of the development and persistence of a specific situation. The main characteristics of longitudinal studies are: 1) an outcome is measured repeatedly within a set of units, 2) longitudinal data are clustered, so repeated measures data are positive correlated within subjects and thus require special statistical techniques for valid analysis and inference, 3) repeated measures obtained from a single subject allow to capture

within-subject patterns of change, 4) the number of repeated observations and their time points can vary widely from one subject to another, therefore having unbalanced longitudinal designs, and 5) missing data that arise when subjects drop out of the study can lead to biased estimates when the probability of missingness is associated with the outcomes.

3.2.2 Sources of variability in longitudinal data

Between-subject variability

Between-subject variation reflects the variation of the mean of each measure from the population mean. In any longitudinal study, some individuals consistently respond higher than the population average, while others tend to respond below the referred average. Thereby, this variability is based on the fact that each studied subject has an underlying behavior (due to genetic, environmental or social factors), and its natural trend can be derived from all repeated measurements taken on.

Within-subject variability or residual variability

It measures the variability in response within the same subject, being sometimes referred as *inherent biological variability*. This variation is based on the variability of scores around the subject's true and unobservable score, because a random measurement error is associated to each biological outcome. Consequently, within-subject variability reflects the part of variance that can not be explained by some predictor variable, given the random effects.

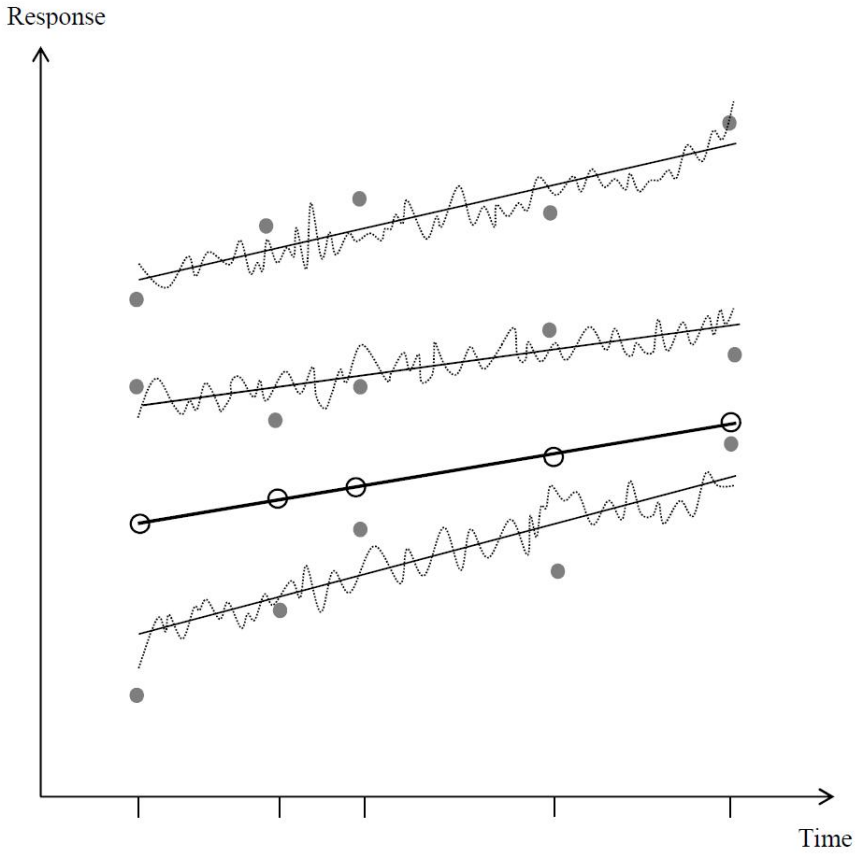

Figure 3.1. Conceptual representation of variation sources in case of balanced longitudinal design with three subjects and five time points: 1) the thick solid line represents the average evolution, 2) the dotted lines represent the pattern of error free responses for the subject over time, 3) the grey points are the observations of these responses, which are subject to measurement error, and 4) the thin solid lines are the trend line for each subject.

3.2.3 The linear mixed effects model

Methodology for analysis of serial data over time

The main goal of linear mixed effects models is to account for the special features of serial evaluations of clinical parameters over time, being able to stablish a plausible model in order to describe the particular evolution of each subject included in a longitudinal design. The particular features of these models are that they can work with unbalanced datasets (unequal number of follow-up measurements between subjects and varying times between repeated measurements of each subject), and that they explicitly take into account that measurements from the same patient may be more correlated than measurements from different patients.

The intuitive idea behind these models is described by Figure 3.2, where it is shown that a linear mixed effects models must have 2 parts in order to describe separately the trend of each subject:

1. Fixed-effects part (solid line in Figure 3.2):

   Describes the average evolution in time of a specific variable under study, where this average is taken overall from the subjects in the sample at hand and is an estimate of the evolution of the longitudinal covariates in the target population. Fixed effects assume that observations are independent.

2. Random-effects part (dashed lines in Figure 3.2):

   Describes the particular evolution in time for each of the subjects under study, taking into account the data correlation within subjects.

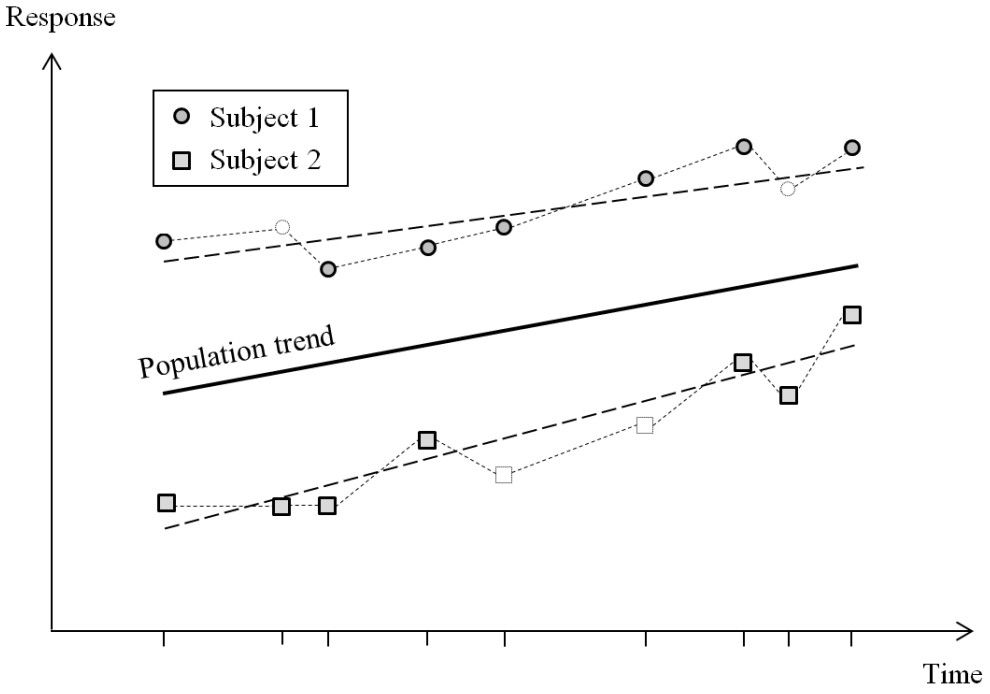

Figure 3.2. Intuitive representation of a linear mixed effects model with two subjects. The points represent hypothetical longitudinal responses for two subjects, and both of them have incomplete data due to drop-out. The dashed lines are the respective subject-specific evolutions, while the solid line defines the population trend.

Generalization of the linear mixed model

Let consider a general longitudinal design with n subjects so that each of them has a different number of n_i repeated measurements of the response variable at different time points. Let denote y_{ij} the response variable on the $i-$th subject, $i = 1, \ldots, n$, observed at time point t_{ij}, $j = 1, \ldots, n_i$, where the outcome is linearly related to a set of p explanatory covariates and q random effects ($q \leq p + 1$). Thus, the set of repeated outcomes for the $i-$th subject can be expressed as the n_i-dimensional vector, $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^{\mathrm{T}}$.

Assuming that the longitudinal outcomes are normally distributed, the general linear mixed model uses the following matrix and vector notation for the $i-$th individual (Laird and Ware, 1982; Verbeke and Molenberghs, 2000):

$$\begin{cases} \mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \\ \mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, \mathbf{D}) \\ \boldsymbol{\varepsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}) \end{cases} \tag{3.1}$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are respectively the $n_i \times (p + 1)$ and $n_i \times q$ design matrices for fixed and random effects, $\boldsymbol{\beta}$ denotes the $(p + 1)-$dimensional vector for the $p + 1$ unknown fixed effects, $\mathbf{b}_i$ is the $q-$dimensional vector for the considered random effects in the model, $\mathbf{D}$ is the $q \times q$ covariance matrix for random effects and $\boldsymbol{\varepsilon}_i$ is the random error n_i-dimensional vector, where σ^2 represents the within-subject variation (constant across the subjects). In addition, random effects, $\{\mathbf{b}_1,\ \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ are assumed to be independent of error terms, i.e., $\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_n\}$.

The above linear mixed model expression can be reformulated in order to describe the subject-specific evolutions at any time t:

$$\begin{cases} y_i(t) = \mathbf{x}_i^{\mathrm{T}}(t)\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}(t)\mathbf{b}_i + \varepsilon_i(t) \\ \mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, \mathbf{D}) \\ \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2) \end{cases} \tag{3.2}$$

where $\mathbf{x}_i^{\mathrm{T}}(t)$ and $\mathbf{z}_i^{\mathrm{T}}(t)$ denote row vectors of the design matrices for the fixed and random effects, respectively.

The interpretation of the $p + 1$ fixed effects coefficients, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$, is exactly the same as in a simple linear regression model, while the coefficients of the random effects $q-$dimensional vector for subject i, $\mathbf{b}_i$, are interpreted in terms of how a subset of the regression parameters for the $i-$th subject deviates from those in the population.

Estimation of linear mixed model parameters

Under the assumption that $\mathbf{b}_i$ and $\boldsymbol{\varepsilon}_i$ are independently distributed as multivariate normal, the estimation of the parameters is based on maximum likelihood inference. In particular, the marginal density of the observed data for the $i-$th subject is given by

$$p(\mathbf{y}_i) = \int_{\mathbf{b}_i} p(\mathbf{y}_i|\mathbf{b}_i)p(\mathbf{b}_i)\mathrm{d}\mathbf{b}_i \tag{3.3}$$

and the above integral leads to the closed-form solution $\mathbf{y}_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i)$, where $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^{\mathrm{T}} + \sigma^2\mathbf{I}_{n_i}$. The variance-covariance matrix can be estimated using the theory of restricted maximum likelihood, REML (Harville, 1977), which is usually preferred to ML because it produces estimators with less bias (Schabenberger and Pierce, 2002). The idea behind REML is to separate the part of the data used in the estimation of $\mathbf{V}_i$ from the part used in the estimation of $\boldsymbol{\beta}$. In general, $\hat{\mathbf{V}}_i$ can not be written in a closed form, so an iterative algorithm is necessary (e.g. Lindstrom and Bates, 1988).

If we know the REML estimation of $\mathbf{V}_i$, the coefficients of the fixed effects vector are estimated by the usual generalized least squares estimator

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} \mathbf{y}_i \ . \tag{3.4}$$

Standard error for the fixed effects coefficients can be obtained by calculating the variance of the above estimator:

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i \right)^{-1} \ . \tag{3.5}$$

Finally, the prediction of the subject-specific effects, $\mathbf{b}_i$, are obtained by the namely best linear unbiased predictors, BLUP's:

$$\hat{\mathbf{b}}_i = \hat{\mathbf{D}}\mathbf{Z}_i^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}) \tag{3.6}$$

where $\hat{\mathbf{D}}$ and $\hat{\mathbf{V}}_i$ are the REML estimators.

Dependence and correlation

The mean and variance of y_{ij} are represented by $\mathrm{E}(y_{ij}) = \mu_{ij}$ and $\mathrm{Var}(y_{ij}) = \mathrm{E}\{(\mathbf{y}_{ij} - \boldsymbol{\mu}_{ij})^2\} = \nu_{ij} = \sigma_j^2$. Therefore, the dependence among the $i-$th subject responses at two different occasions, say y_{ij} and y_{ik}, can be denoted by the covariance expression, $\mathrm{Cov}(y_{ij}, y_{ik}) = E\{(y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})\} = \nu_{ijk} = \sigma_{ij}$. The $n_i \times n_i$ covariance matrix for the $i-$th subject is:

$$\mathrm{V}_i = \mathrm{Var}(\mathbf{y}_i) = \mathrm{Var} \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n_i} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n_i1} & \sigma_{n_i2} & \cdots & \sigma_{n_i}^2 \end{pmatrix} \tag{3.7}$$

Due to the fact that the magnitude of the covariance is somewhat difficult to interpret without comparing it to the underlying variability of the variables, the correlation term between y_{ij} and y_{ik}, ρ_{jk} is widely used: $\rho_{jk} = E\{(y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})\}/(\sigma_j\sigma_k)$, being σ_j and σ_k the standard deviations

of y_{ij} and y_{ik}, respectively. We can also define the $n_i \times n_i$ correlation matrix for the $i-$th subject:

$$\mathrm{R}_i = \mathrm{Corr}(\mathbf{y}_i) = \mathrm{Corr}\begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{pmatrix} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n_i} \\ \rho_{21} & 1 & \cdots & \rho_{2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n_i1} & \rho_{n_i2} & \cdots & 1 \end{pmatrix} \tag{3.8}$$

With longitudinal data, repeated observations on the same individual are not independent, and the variance of repeated measurements is not usually constant. Consequently: 1) the heterogenity of variance over time is accounted by allowing the elements on the main diagonal of the covariance matrix to differ, and 2) dependence between responses leads that the off-diagonal elements of the covariance matrix are usually non-zero, while these same elements are positive in the correlation matrix.

Modeling covariance structure refers to representing V_i in expression (3.7) as a function of a relatively small number of parameters. Functional specification of the covariance structure for the linear mixed model is done through $\sigma^2\mathbf{I}_n$ and $\mathbf{D}$ matrice. For simplicity, repeated measurements are equally spaced assumed, so we may define the following basic structures: 1) Simple structure specifies that the observations are independent, even on the same patient, 2) Compound symmetric structure, which specifies that observations on the same patient have homogeneous covariance homogeneous variance, 3) Autoregressive (order 1) covariance structure, which specifies homogeneous variance and also that covariances between observations on the subject patient are not equal, but decrease toward zero with increasing lag, 4)Autoregressive plus random effects structure specifies that covariance between observations on the same patient comes from two sources and 5) The unstructured covariance specifies no patterns in the covariance matrix.

3.3 Survival Data Analysis

3.3.1 General features of survival data

The term *survival analysis* is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of a specific event of interest, usually designed by $\mathcal{E}$. This time is called *survival time* or *event time*.

Bearing the above into account, what makes survival data special is that the responses are times and thus are not measured in the same way as other variables. In practice, this fact has two important consequences: 1) the distribution of survival times is often highly left-skewed, and 2) the only information we can have about some subjects is that they have not yet experienced the event $\mathcal{E}$ at the last time point of follow-up, so these are termed censored or incomplete observations (we do not know when these remaining subjects will experience the event). Considering these two special features, standard statistical methods can not be applied to survival data.

Although there are various categories of censoring, **the present work has only focused in right-censoring mechanism**, which occurs when the subject has not yet experienced the event of interest at the time when the follow-up period ends. Consequently, all that is known about the true survival time, t^*, is that it exceeds the observed survival time, t, at the end study. Furthermore, over all this work **it will be considered that the censoring is non-informative**. In this regard, it will be assumed the following three basic reasons why right-censoring might occur:

- The event of interest has not occurred by the end of the follow-up period (study end).

- A subject is lost to follow-up during the study period due to not concerned causes to the event of interest.
- A subject withdraws from the study (dropout) due to not concerned causes to the event of interest.

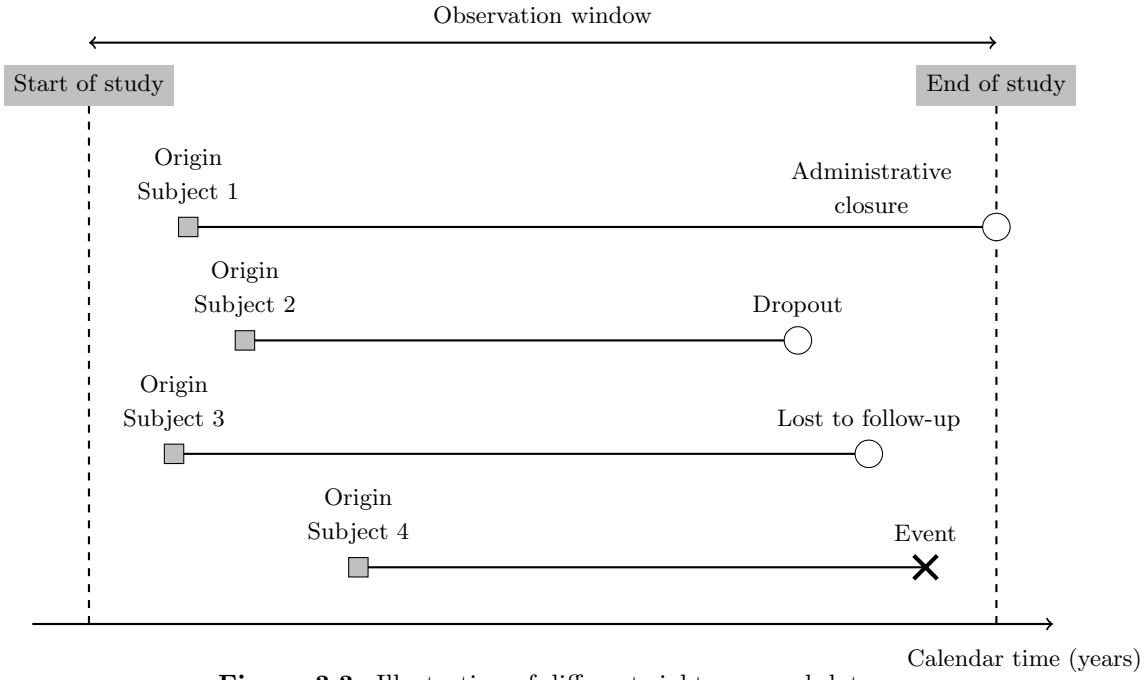

Figure 3.3. Illustration of different right-censored data cases.

3.3.2 Main survival functions

From now on, we assume that T is a non-negative continuous random variable which represents the time until some specified event. Lets assume that its cumulative distribution function is $F(t)$ and its probability density function is $f(t)$. In this situation, two functions are of central interest in survival analysis:

- Survival function, $S(t)$

  It denotes the probability of an individual surviving beyond time t, that is, the probability that the event of interest has not yet occurred before time t. It is defined as

$$S(t) = \Pr(T > t) = 1 - F(t), \text{ for } t \geq 0. \tag{3.9}$$

- Hazard function, $h(t)$

  It represents the rate of occurrence of the event of interest at a given time t,

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(t < T < t + \Delta t | T \geq t)}{\Delta t} \text{ , with } h(t) \geq 0. \tag{3.10}$$

  Therefore:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} \left[\log\{S(t)\}\right] \tag{3.11}$$

3.3.3 Non-parametric analysis of survival data

Kaplan-Meier estimator of the survival function

Let T be a positive random variable which represents the observed survival time, with survival function $S(t)$, $t \geq 0$. Given a random sample of $i = 1, \ldots, n$ individuals, the $i-$th subject provides a potential true survival time, T_i^*, and a potential right-censored time, C_i, so that each individual can be summarized by a couple of values, the observed survival time, $T_i = \min\{T_i^*, C_i\}$ and an event indicator $\delta_i = I(T_i^* \leq C_i)$. In most cases, we can have ties between observed times, so two or more individuals in the dataset share the observed survival time. Therefore, we only observe a number of r ($r < n$) distinct survival times: in the same time there might be more than one event observation, more than one right-censored observation or both type of survival data.

Let consider the order statistics of the r different observed survival times, $T_{(1)} < T_{(2)} < \ldots < T_{(r)}$, for $j = 1, \ldots, r$. Then, if $t_{max} = T_{(r)}$ is the largest observation time, the Kaplan-Meier estimator (Kaplan and Meier, 1958) admits the following general form for all $t < t_{max}$:

$$\hat{S}_{KM}(t) = \begin{cases} 1 & \text{if } t < T_{(1)} \\ \prod_{j:T_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right) & \text{if } t \geq T_{(1)} \end{cases} \tag{3.12}$$

where $n_j = card(R(T_{(j)}))$ is the number of individuals who are at risk at a given observed time $T_{(j)}$, and d_j is the number of individuals who experience the event of interest at $T_{(j)}$. For values of t beyond the largest observation time, t_{max} this estimator is not well defined.

Tests for two or more samples

Comparing two or more populations when when working with censored data can be performed by simply comparing their respective estimated survival curves. Assuming two distinct groups, the comparison is performed by the following non-parametric contrast:

$$\begin{aligned} &H_0 : S_1(t) = S_2(t) \ \ \forall\, t \text{ such that } 0 \leq t \leq \tau \\ &H_1 : S_1(t) \neq S_2(t) \text{ for some } t = \hat{t} \text{ such that } 0 \leq \hat{t} \leq \tau \end{aligned} \tag{3.13}$$

where $[0, \tau]$ is the window observation.

Let be $t_1 < t_2 < ... < t_D$, with $D \leq n$, the ordered time points with any event in the pooled sample. In each time with any event, t_i, we can observe d_{i1} events and R_{i1} subjects at risk in group 1, and d_{i2} events and R_{i2} subjects at risk in group 2. Then, for the $i-$th time it is possible to construct a 2×2 contingency table:

Group	Events	Survivors	Total at risk
1	d_{i1}	$R_{i1} - d_{i1}$	R_{i1}
2	d_{i2}	$R_{i2} - d_{i2}$	R_{i2}
Total	d_i	$R_i - d_i$	R_i

Table 3.1. Explanatory table 2×2 with information on the number of events and the number of individuals at risk, per group and in the overall sample, in the $i-$th global event time.

By considering the differences between the observed event rates in each group, d_{ij} for $j = 1, 2$, and the expected rates, d_i/R_i, tests can be constructed to weight the differences between the two rates by suitably chosen weights, $W_j(t_i)$. The term usually applied to the weight function is $W_j(t_i) = R_{ij}W(t_i)$, so that the statistic test can be expressed as:

$$Z_W(\tau) = \frac{\sum_{i=1}^{D} W(t_i)\left(d_{i1} - R_{i1}\frac{d_i}{R_i}\right)}{\sqrt{\sum_{i=1}^{D} W^2(t_i)\frac{R_{i1}}{R_i}\left(1 - \frac{R_{i1}}{R_i}\right)\left(\frac{R_i - d_i}{R_i - 1}\right)}} \tag{3.14}$$

The evidence against the null hypothesis is summarized by the value $Z_W(\tau)$, so that under H_0 it is verified that $[Z_W(\tau)]^2 \sim \chi_1^2$ when the sample size is large enough. The weights used throughout this work are those referred to the so-called G^ρ family, proposed by Harrington and Fleming (1982):

$$W_{FH}(t_i) = \left[\hat{S}_{KM}(t_{i-1})\right]^\rho \tag{3.15}$$

where ρ is real value. When $\rho = 0$, the same weight is given to all differences between the two estimated survival curves, corresponding to the so called *log-rank* test.

To conclude this section, it must be marked that we would follow the same reasoning as the exposed in case of having more than two populations.

3.3.4 The proportional-hazards Cox model

Implementation of PH Cox model for censored survival data

The celebrated proportional-hazards Cox model (Cox, 1972) allows to model the conditional hazard rate of survival times given certain baseline covariates. It relies on a fundamental assumption, the proportionality of the hazards, implying that the factors investigated have a constant impact on the risk over time. The model provides the conditional hazard function $h_i(t|\mathbf{w}_i)$ at time t of a subject's profile given by a set of p time-independent explanatory covariates (so-called baseline covariates), $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{ip})^{\mathrm{T}}$:

$$h_i(t|\mathbf{w}_i) = h_0(t)\psi(\mathbf{w}_i)\ ,\ t \geq 0 \tag{3.16}$$

where:

$h_0(t)$: is an unspecified and non-negative baseline hazard function, representing the hazard function when $\mathbf{w}_i = \mathbf{0}$.

$\psi(\mathbf{w}_i)$: is a non-negative function which contains the information about the set explanatory time-independent covariates that define the $i-$th subject's profile.

This model is defined as a semiparametric because a parametric form is assumed only for the covariate effect, $\psi(\mathbf{w}_i)$. Among the possible parameterizations of the function ψ, the most used is the one adopted an exponential expression: $\psi(\mathbf{w}_i;\boldsymbol{\gamma}) = \exp(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_i)$, where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)^{\mathrm{T}}$ is the parameter vector:

$$h_i(t|\mathbf{w}_i) = h_0(t)\exp\{\gamma_1 w_{i1} + \gamma_2 w_{i2} + \ldots + \gamma_p w_{ip}\} = h_0(t)\exp\left\{\sum_{k=1}^{p}\gamma_k w_{ik}\right\} \tag{3.17}$$

The Cox model is often called a proportional hazards (PH) model because, if we look at two individuals i and i' with respective covariate values $\mathbf{w}_i$ and $\mathbf{w}_{i'}$, the ratio between their hazard rates, *HR*, is constant, so their hazard rates are proportional to each other and do not depend on time:

$$HR = \frac{h(t|\mathbf{w}_i)}{h(t|\mathbf{w}_{i'})} = \frac{h_0(t)\exp\left\{\sum_{k=1}^{p}\gamma_k w_{ik}\right\}}{h_0(t)\exp\left\{\sum_{k=1}^{p}\gamma_k w_{i'k}\right\}} = \exp\left\{\sum_{k=1}^{p}\gamma_k(w_{ik}-w_{i'k})\right\} \tag{3.18}$$

In the absence of any information for data distribution, the Cox model is a robust option for different reasons: 1) The exponential part ensures that the estimated hazards are non-negative, 2) We can estimate the γ_k's in the exponential part of the model, and 3) It is preferred over the logistic model when survival time information is available and there is censoring.

Estimation of PH Cox model parameters

When the baseline hazard function is completely unspecified and the form of the function $\psi(\cdot)$ is given, inference can be based on the namely partial likelihood function (Cox, 1975), which does not require specification of $h_0(\cdot)$. Assuming the presence of ties, information about parameters of the model can be obtained from the relative orderings (i.e., ranks) of the survival times. Let consider the d different ordered event times, $t^*_{(1)} < t^*_{(2)} < \ldots < t^*_{(d)}$, and denote $\mathcal{R}_j = \mathcal{R}(t_{(j)}) = \{i : T_i \geq t_{(j)}\}$, as the set of individuals who are "at risk" for experiencing the event at time $t_{(j)}$, $j = 1, \ldots, d$ and $\mathcal{R}(T_i)$ as the risk set at the event time of the $i-$th subject, $i = 1, \ldots, n$. Intuitively, the partial likelihood function is a product over the set of observed event times of the conditional probabilities of seeing the observed events, given the set of individuals at risk at those times:

$$\mathcal{L}_p \propto \prod_{i=1}^{n}\frac{\exp\{\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_i\}}{\sum_{j\in\mathcal{R}(T_i)}\exp\{\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_j\}} \Rightarrow \log\mathcal{L}_p \propto \sum_{i=1}^{n}\delta_i\left\{\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_i - \log\left[\sum_{j\in\mathcal{R}(T_i)}\exp(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_j)\right]\right\} \tag{3.19}$$

In particular, the maximum partial likelihood estimators are found by solving the partial log-likelihood score equations:

$$\frac{\partial\log\mathcal{L}_p}{\partial\boldsymbol{\gamma}^{\mathrm{T}}} = \sum_{i=1}^{n}\delta_i\left\{\mathbf{w}_i - \frac{\sum_{j\in\mathcal{R}(T_i)}\mathbf{w}_j\exp(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_j)}{\sum_{j\in\mathcal{R}(T_i)}\exp(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_j)}\right\} = \mathbf{0} \tag{3.20}$$

It can be demonstrated that the maximum likelihood obtained from the maximization, $\hat{\boldsymbol{\gamma}}$, is asymptotically normal an unbiased, efficient and asymptotically normal distributed, so if $n \to \infty$ then $\hat{\boldsymbol{\gamma}} \sim N\big(\boldsymbol{\gamma},\ \{E[\mathcal{I}(\boldsymbol{\gamma})]\}^{-1}\big)$, where $\mathcal{I}(\boldsymbol{\gamma})$ is the Fisher matrix information.

In the presence of tied survival times, there are three widely used methods to treat ties between event times in the Cox proportional hazards model: The Breslow approximation (Breslow, 1974), the Efron approximation (Efron, 1977) and the exact method (Kalbfleisch and Prentice, 2002).

3.3.5 The extended Cox model with time-dependent covariates

So far, we have been considering the Cox PH model, where the $i-$th subject baseline covariates $\mathbf{w}_i$, $i = 1, \ldots, n$ are measured at study entry ($t = 0$). However, since survival data occur over time, important covariates that we wish to consider may also change within the observation period. We refer to these as time-dependent covariates, $\mathbf{y}_i(t)$.

Types of time-dependent covariates

Let $\mathbf{y}_i(t)$ denote the covariate vector at time t for the $i-$th subject, and let be $\mathcal{Y}_i(t) = \{y_i(s), 0 \leq s \leq t\}$ the associated covariate history up to time point t. It is very important to differentiate between two different types of time-dependent covariates:

1. Exogenous covariates

   The formal definition of exogenous covariates (Kalbfleisch and Prentice, 2002) requires such covariates to satisfy the relation $\forall s, t$ such that $0 \leq s < t$ and $ds \to 0$:

   $$\Pr(s \leq T_i^* < s + ds | T_i^* \geq s, \mathcal{Y}_i(s)) = \Pr(s \leq T_i^* < s + ds | T_i^* \geq s, \mathcal{Y}_i(t)) \tag{3.21}$$

   The above expression formalizes the idea that $y_i(.)$ is associated with the rate of events over time, but its future path to any time $t > s$ will not be affected by the occurrence of event at time s. Thus, an external covariate is one whose path is completely predictable, i.e., its value at any time t can be known infinitesimally before t (e.g., the time of the day).

2. Endogenous covariates

   An endogenous time-dependent covariate is one where the change of the covariate over time is related to the behavior of the individual, so it is measured with error and we can not predict the future path of the covariate value. In this situation, the complete history is not available, but measurements are only available at specific time-points.

Implementation of the extended Cox model

The local nature of the proportional hazards model reformulates itself easily to extensions that allows for covariates that change over time. The basic idea lied in the fact of considering a counting process model, where a subject contributes to the risk set for an event as long as an individual is under observation at the time the event occurs and shares the same baseline hazards function (Andersen and Gill, 1982; Therneau and Grambsch, 2000). In counting process notation, the event process is written as $\{N_i(t), R_i(t)\}$, with $N_i(t)$ denoting the number of events for subject i at time point t and $R_i(t)$ is at-risk indicator for the mentioned subject at time t. Let be:

$\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{ip})^{\mathrm{T}}$, vector of baseline covariates associated with the hazard of each subject

$\mathbf{y}_i(t)$, covariate vector for subject i at time t

$\mathcal{Y}_i(t) = \{y_i(s), 0 \leq s \leq t\}$, covariate history for subject i up to time t

Then, a namely counting process model with instantaneous rate of occurrence of the event of interest can be formulated as

$$h_i(t|\mathcal{Y}_i(t), \mathbf{w}_i) = h_0(t) R_i(t) \exp\{\boldsymbol{\gamma}^{\mathrm{T}} \mathbf{w}_i + \alpha y_i(t)\} \tag{3.22}$$

where the regression coefficients $\boldsymbol{\gamma}$ and the parameter α has analogous interpretation as the coefficients in the PH Cox model. In this case, the interpretation refers to a subject's particular time point, t. The estimation of $\boldsymbol{\gamma}$ and α is again based on the partial log-likelihood function:

$$\mathcal{L}_p(\boldsymbol{\gamma}) \propto \sum_{i=1}^{n} \int_0^{\infty} \left(R_i(t) \exp\{\boldsymbol{\gamma}^{\mathrm{T}} \mathbf{w}_i + \alpha y_i(t)\} - \log \left[\sum_{j \in \mathcal{R}(T_i)} R_j(t) \exp\{\boldsymbol{\gamma}^{\mathrm{T}} \mathbf{w}_i + \alpha y_i(t)\} \right] \right) \mathrm{d}N_i(t) \tag{3.23}$$

Considerations on the extended Cox model

In the case of endogenous variables (the most common in clinical research) we do not know all the covariate history. To try to solve this lack of information, time-dependent outcomes are assumed to change value at follow-up time points, while remaining constant between these timings. In this regard, Figure 3.4 shows the particular way the Cox model handles time-dependent covariates under the counting process formulation:

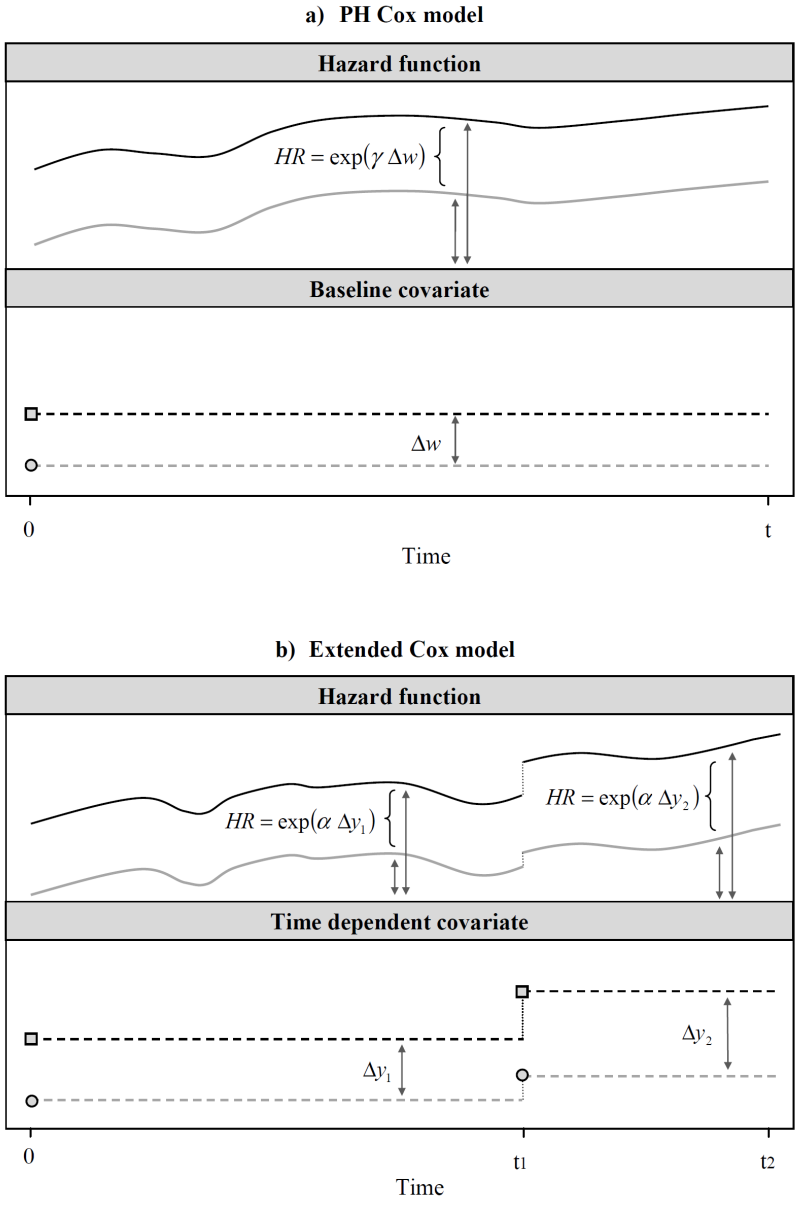

Figure 3.4. Comparison on the treatment of baseline covariates in the PH Cox model (top panels) and time-dependent covariates in the extended Cox model (bottom panels). Working with time-dependent covariates, the PH assumption is only valid between consecutive time points.

However, it is not reasonable to assume that an endogenous covariate remains constant between measurement points, specially when these points can be months or even years apart. In particular, in the case of biomarkers this approach generally leads to biased estimations and standard errors (Prentice, 1982).

3.4 Joint Modeling framework

3.4.1 The classical Joint Modeling approach

The longitudinal submodel

To account for the fact that the longitudinal marker is an endogenous time-dependent covariate measured with error (Kalbfleisch and Prentice, 2002) , it is assumed that the risk for an event depends on the true and unobserved value of the endogenous variable at time t, denoted by $m_i(t)$.

Therefore, it must be estimated $m_i(t)$ in order to successfully reconstruct the complete longitudinal history $\mathcal{M}_i(t)$. For this purpose, we utilize all the available measurements on each subject $\{y_i(t_{ij}),\ \text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, n_i\}$ and postulate a suitable mixed effects model. We will focus on normal data, describing the true subject-specific evolutions by a linear mixed effects model:

$$\begin{cases} y_i(t) = m_i(t) + \varepsilon_i(t) = \mathbf{x}_i^{\mathrm{T}}(t)\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}(t)\mathbf{b}_i + \varepsilon_i(t) \\ \mathbf{b}_i \sim \mathcal{N}_2(\mathbf{0}, \mathbf{D}) \\ \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2) \end{cases} \tag{3.24}$$

The survival submodel

In order to quantify the effect of the true outcome $m_i(t)$ on the risk for an event at specific time t, we use a relative risk model of the form (Therneau and Grambsch, 2000):

$$h_i(t|\mathcal{M}_i(t), \mathbf{w}_i) = h_0(t) \exp\{\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_i + \alpha m_i(t)\}\ ,\ t > 0 \tag{3.25}$$

where $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s \leq t\}$ denotes the history of the true unobserved longitudinal process for subject i up to time point t. The parameter α quantifies the grade of association between the true marker terms and the risk for an event.

In standard survival analysis, the baseline risk function $h_0(\cdot)$ is typically left completely unspecified (Cox, 1972; Andersen and Gill, 1982). However, within the joint modeling framework (Hsieh et al., 2006) noted that leaving this function completely unspecified leads to an underestimation of the standard errors of the parameter estimates, so it is necessary to explicitly define $h_0(\cdot)$. Although we could use the hazard function of a standard survival distribution (e.g. Weibull or Gamma), we finally opted for a more flexible solution such a piecewise-constant model:

$$h_0(t) = \sum_{q=1}^{Q} \xi_q I(\nu_{q-1} < t \leq \nu_q) \tag{3.26}$$

where $0 = \nu_0 < \nu_1 < \ldots < \nu_Q$ denotes a split of the time scal, with ν_Q being the largest observed time, and ξ_Q denotes the value of the hazard in the interval $(\nu_{q-1}, \nu_q]$.

The joint model formulation

On the basis of the expressed considerations, the true and unobserved outcome at a specific time point t can be modeled by joining the two above approaches (Rizopoulos, 2012b):

$$\begin{cases} y_i(t) = m_i(t) + \varepsilon_i(t) = \mathbf{x}_i^{\mathrm{T}}(t)\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}(t)\mathbf{b}_i + \varepsilon_i(t) \\ h_i(t|\mathcal{M}_i(t), \mathbf{w}_i) = h_0(t)R_i(t)\exp\{\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_i + \alpha m_i(t)\} \end{cases} \tag{3.27}$$

Particularly, the hazard at age t for the $i-$th individual, with a true longitudinal profile $\mathcal{M}_i(t)$ up to time t, can be expressed as follows:

$$h_i(t|\mathcal{M}_i(t), \mathbf{w}_i) = h_0(t)R_i(t)\exp\left[\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_i + \alpha\{\mathbf{x}_i^{\mathrm{T}}(t)\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}(t)\mathbf{b}_i\}\right] \tag{3.28}$$

3.4.2 Estimation of parameters in joint modeling

The estimation method proposed for joint models in this work is maximum likelihood (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Hsieh et al., 2006), based on the maximization of the log-likelihood corresponding to the joint distribution of the observed time-to-event and longitudinal outcomes $(T_i, \delta_i, \mathbf{y}_i)$. To define this joint distribution we will assume that the vector of time-independent random effects bi underlies both the longitudinal and survival processes. This means that these random effects account for both the association between the longitudinal and event outcomes, and the correlation between the repeated measurements in the longitudinal process (conditional independence). Assuming non differential measurement error, we have:

$$p(T_i, \delta_i, \mathbf{y}_i|\mathbf{b}_i; \boldsymbol{\theta}) = p(T_i, \delta_i|\mathbf{b}_i; \boldsymbol{\theta})p(\mathbf{y}_i|\mathbf{b}_i; \boldsymbol{\theta}) \tag{3.29}$$

$$p(\mathbf{y}_i|\mathbf{b}_i; \boldsymbol{\theta}) = \prod_j p\{y_i(t_{ij})|\mathbf{b}_i; \boldsymbol{\theta}\} \tag{3.30}$$

where $\mathbf{y}_i$ is the n_i-vector of longitudinal responses of the $i-$th subject, $p(\cdot)$ an appropriate probability density function, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_t^{\mathrm{T}}, \boldsymbol{\theta}_y^{\mathrm{T}}, \boldsymbol{\theta}_b^{\mathrm{T}})^{\mathrm{T}}$ the full parameter vector, with $\boldsymbol{\theta}_t$ denoting the parameters for the event time outcome, $\boldsymbol{\theta}_y$ the parameters for the longitudinal outcomes and $\boldsymbol{\theta}_b$ the unique parameters of the random-effects covariance matrix. In addition, we assume that given the observed history, the censoring mechanism and the visiting process are independent of the true event times and future longitudinal measurements. By visiting process we define the stochastic mechanism that generates the time points at which longitudinal measurements are collected (Lipsitz et al., 2002), whereas for any time point t, we define as observed history all available information for the longitudinal process prior to t.

The joint log-likelihood contribution for the $i-$th subject can be formulated as:

$$\begin{aligned} \log p(T_i, \delta_i, \mathbf{y}_i; \boldsymbol{\theta}) &= \log \int_{\mathbf{b}_i} p(T_i, \delta_i, \mathbf{y}_i; \boldsymbol{\theta})\mathrm{d}\mathbf{b}_i = \\ &= \log \int_{\mathbf{b}_i} p(T_i, \delta_i|\mathbf{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) \left[\prod_j p\{y_i(t_{ij})|\mathbf{b}_i; \boldsymbol{\theta}_y\}\right] p(\mathbf{b}_i; \boldsymbol{\theta}_b)\mathrm{d}\mathbf{b}_i \end{aligned} \tag{3.31}$$

where the conditional density for the survival part is written as:

$$p(T_i, \delta_i|\mathbf{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) = \{h_i(T_i|\mathcal{M}_i(T_i); \boldsymbol{\theta}_t, \boldsymbol{\beta})\}^{\delta_i} \mathcal{S}_i(T_i|\mathcal{M}_i(T_i); \boldsymbol{\theta}_t, \boldsymbol{\beta}) =$$
$$= \left[h_0(T_i) \exp\{\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_i + \alpha m_i(t)\}\right]^{\delta_i} \exp\left(- \int_0^{T_i} h_0(s)\{\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{w}_i + \alpha m_i(s)\}\mathrm{d}s \right) \tag{3.32}$$

In equation 3.31, $p\{y_i(t_{ij})|\mathbf{b}_i; \boldsymbol{\theta}_y\}$ is the univariate normal density for the longitudinal responses, and $p(\mathbf{b}_i; \boldsymbol{\theta}_b)$ is the multivariate normal density for the random effects.

The maximization of the log-likelihood function, $\log \mathcal{L}(\boldsymbol{\theta}) = \sum_i \log p(T_i, \delta_i, \mathbf{y}_i; \boldsymbol{\theta})$ can be achieved using standard algorithms. In the joint modeling literature, the Expectation-Maximization algorithm, EM (Dempster et al., 1977) has been traditionally used, which is a very general iterative algorithm for incomplete-data problems. With this regard, (Rizopoulos et al., 2009) noted that the score vector corresponding to $\log \mathcal{L}(\boldsymbol{\theta})$ is the key function required in the EM algorithm. The score vector can be rewritten in the form:

$$S(\boldsymbol{\theta}) = \frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} = \sum_i \int_{\mathbf{b}_i} A(\boldsymbol{\theta}, \mathrm{b}_i) p(\mathbf{b}_i|T_i, \delta_i, \mathbf{y}_i; \boldsymbol{\theta}) \mathrm{d}\mathbf{b}_i \tag{3.33}$$

where $A(\cdot)$ denotes the complete data score vector given by

$$A(\theta, \mathbf{b}_i) = \partial\{\log p(T_i, \delta_i|\mathbf{b}_i; \boldsymbol{\theta}) + \log p(\mathbf{y}_i|\mathbf{b}_i; \boldsymbol{\theta}) + \log p(\mathbf{b}_i; \boldsymbol{\theta})\}/\partial \boldsymbol{\theta}^{\mathrm{T}}$$

When the score equations corresponding to (3.33) are solved with respect to $\boldsymbol{\theta}$, with $p(\mathbf{b}_i|Ti, \delta_i, \mathbf{y}_i; \boldsymbol{\theta})$ fixed at the $\boldsymbol{\theta}$ of the previous iteration, then this corresponds to an EM algorithm. In contrast, if the score equations are solved with respect to $\boldsymbol{\theta}$ considering $p(\mathbf{b}_i|Ti, \delta_i, \mathbf{y}_i; \boldsymbol{\theta})$ also a function of $\boldsymbol{\theta}$, then this corresponds to a direct maximization of the observed data log-likelihood, $\log \mathcal{L}_p(\boldsymbol{\theta})$

Standard errors for the parameter estimates can be based on the estimated observed information matrix, i.e.,

$$\hat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) = \left\{\mathcal{I}(\hat{\boldsymbol{\theta}})\right\}^{-1} \ , \mathcal{I}(\hat{\boldsymbol{\theta}}) = -\sum_{i=1}^{n} \frac{\partial S_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{3.34}$$

3.4.3 Residual analysis of joint models

The standard tools to assess model assumptions are residual plots, which can be used to determine the adequacy of the fitted joint model and can also indicate the presence of outliers. Some of the most notable works for linear mixed models diagnostics are found in Santos Nobre and da Motta Singer (2007), whereas reference papers focused on residuals for survival models were presented by Harrell (2001) and Therneau and Grambsch (2000). However, getting residuals based on the fitted joint model and the observed data is not straightforward, since they can be subject to a informative not at random dropout mechanism. In order to avoid this difficulty, Rizopoulos et al. (2010) suggested to impute longitudinal responses under the complete data model, thereby working from a Bayesian perspective (Little and Rubin, 2002).

With regard to the longitudinal submodel, the fitted model is often checked by using subject-specific (conditional) standardized residuals, $r_i^{(yss)}(t_{ij}) = \{y_i(t_{ij}) - \mathbf{x}_i^{\mathrm{T}}(t_{ij})\hat{\boldsymbol{\beta}} - \mathbf{z}_i^{\mathrm{T}}(t_{ij})\hat{\mathbf{b}}_i\}/\hat{\sigma}$, and marginal

(population averaged) standardized residuals, $\mathbf{r}_i^{(ysm)} = \hat{\mathbf{V}}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ denote the maximum likelihood estimates under linear mixed model, $\hat{\mathbf{b}}_i$ are the empirical Bayes estimates for the random effects, and $\hat{\mathbf{V}}_i$ is the estimated marginal covariance matrix of $\mathbf{y}_i$.

For the survival submodel, the validation strategy is based on the martingale residuals, $r_i^{(tm)}(t_{ij})$, and the Cox-Snell residuals, $r_i^{(tcs)}$. Martingale residuals focus on the counting process formulation, and are obtained both at each subject's time point, $\{t_{ij},\ i = 1,\ldots,n,\ j = 1,\ldots,n_i\}$ where the longitudinal outcome was measured, and at every observed survival time, $T_i,\ i = 1,\ldots,n$. If we denote by $N_i(t_{ij})$ the counting process for i-th subject at time t_{ij}, martingale residuals are written as $r_i^{(tm)}(t_{ij}) = N_i(t_{ij}) - \int_0^{t_{ij}} R_i(s)\hat{h}_0(s)\exp\{\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{w}_i + \hat{\alpha}\hat{m}_i(s)\}\mathrm{d}s$, and Cox-Snell residuals as $r_i^{(tcs)} = \int_0^{T_i} \hat{h}_0(s)\exp\{\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{w}_i + \hat{\alpha}\hat{m}_i(s)\}\mathrm{d}s$.

The problem in using the above defined residuals for inspecting the fit of joint models is that their reference distribution is not directly evident. Complications arise due to the non random dropout in the longitudinal process caused by the occurrence of events. That is, the observed data, upon which the residuals are calculated, are not a random sample of the target population. To clarify this, we define for each subject the observed and missing part of the longitudinal response vector. The observed part $\mathbf{y}_i^o = \{y_i(t_{ij}) : t_{ij} < T_i,\ j = 1,\ldots,n_i\}$ contains all observed longitudinal measurements of the $i-$th subject before the observed event time, whereas the missing part $\mathbf{y}_i^m = \{y_i(t_{ij}) : t_{ij} \geq T_i,\ j = 1,\ldots,n_i'\}$ contains the longitudinal measurements that would have been taken until the end of the study, had the event not occurred. Under these definitions, it can be derived the dropout mechanism, which is the conditional distribution of the time-to-dropout given the complete vector of longitudinal responses $(\mathbf{y}_i^o, \mathbf{y}_i^m)$,

$$p(T_i^* \mid \mathbf{y}_i^o, \mathbf{y}_i^m; \boldsymbol{\theta}) = \int_{\mathbf{b}_i} p(T_i^*|\mathbf{b}_i; \boldsymbol{\theta}) p(\mathbf{b}_i \mid \mathbf{y}_i^o, \mathbf{y}_i^m; \boldsymbol{\theta})\mathrm{d}\mathbf{b}_i \tag{3.35}$$

which still depends on $\mathbf{y}_i^m$ through the posterior distribution $p(\mathbf{b}_i|\mathbf{y}_i^o, \mathbf{y}_i^m; \boldsymbol{\theta})$. It is this feature of joint models that complicates inspection of residual plots, because a potential systematic behavior is not necessarily indicative of a model misfit. Thus, conclusions from common residual plots in the joint model framework should be drawn with extreme caution.

3.4.4 Predicted survival in joint models

Once the model has been validated, a powerful feature is to derive satisfactory results in terms of survival predictions. Thus, considering the sample $\mathcal{D}_n = \{T_i, \delta_i, y_i;\ i = 1,\ldots,n\}$ on which the joint model was fitted, the goal consists of predicting conditional probability of surviving time for a new subject i that provides a set of longitudinal measurements, $\mathcal{Y}_i(t) = \{y_i(s);\ 0 \leq s < t\}$ and a vector of baseline covariates, $\mathbf{w}_i$. Flexibility provided by joint modeling approach is in line with a growing trend towards personalized medicine (Garre et al., 2008; Proust-Lima and Taylor, 2009; Rizopoulos, 2011). In particular, the real challenge focuses of being able to estimate these probabilities not only at each one of the time points measurements, but also at a generic time $u > t$ given survival up to t

$$\pi_i(u|t) = \Pr(T_i^* \geq u \mid T_i^* > t, \mathcal{Y}_i(t), \mathbf{w}_i, \mathcal{D}_n; \boldsymbol{\theta}^*) \tag{3.36}$$

where $\boldsymbol{\theta}^*$ denotes the true parameter values.

This approach therefore allows to obtain the so called survival dynamic predictions for the $i-$th subject, arising from its survival curve updating on the basis of any new longitudinal information

subsequently collected. Hence, as new information at time $t' > t$ joins existing longitudinal measurements, one can update the estimated survival curve $\pi_i(u \mid t)$ to $\pi_i(u \mid t')$, and therefore proceed in a *time dynamic* manner.

The estimation of the subject-specific conditional survival probabilities takes full advantage of the conditional independence used to define the joint model. Using a Bayesian formulation (Proust-Lima and Taylor, 2009; Rizopoulos, 2011), the problem can be written as:

$$\Pr(T_i^* \geq u \mid T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_n) = \int_{\boldsymbol{\theta}} \Pr(T_i^* \geq u \mid T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}_n) \mathrm{d}\boldsymbol{\theta} \tag{3.37}$$

The first part of the above integrand is given by

$$\Pr(T_i^* \geq u \mid T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) = \int_{\mathbf{b}_i} \frac{S_i\{u \mid \mathcal{M}_i(u, \mathbf{b}_i, \boldsymbol{\theta}); \boldsymbol{\theta}\}}{S_i\{t \mid \mathcal{M}_i(t, \mathbf{b}_i, \boldsymbol{\theta}); \boldsymbol{\theta}\}} \, p(\mathbf{b}_i \mid T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) \mathrm{d}\mathbf{b}_i \tag{3.38}$$

where $S_i(\cdot)$ denotes the survival function, and furthermore it has been explicitly noted that the true longitudinal history $\mathcal{M}_i(\cdot)$ is a function of both the random effects and the parameters. For the second part of equation (3.37), it is assumed that the sample size n is sufficiently large, such that $\{\boldsymbol{\theta}, \mathcal{D}_n\}$ can be well approximated by $\mathcal{N}\{\hat{\boldsymbol{\theta}}, \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\}$.

By combining 3.37, 3.38 and $\{\boldsymbol{\theta}, \mathcal{D}_n\} \sim \mathcal{N}\{\hat{\boldsymbol{\theta}}, \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\}$, it can be derived a Monte Carlo estimate of $\pi_i(u \mid t)$ using the following simulation scheme:

1) Draw $\boldsymbol{\theta}^{(l)} \sim \mathcal{N}\{\hat{\boldsymbol{\theta}}, \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\}$

2) Draw $\mathbf{b}_i^{(l)}$

3) Compute $\pi_i^{(l)}(u \mid t) = S_i\{u \mid \mathcal{M}_i(u, \mathbf{b}_i^{(l)}, \boldsymbol{\theta}^{(l)}); \boldsymbol{\theta}^{(l)}\} / S_i\{t \mid \mathcal{M}_i(t, \mathbf{b}_i^{(l)}, \boldsymbol{\theta}^{(l)}); \boldsymbol{\theta}^{(l)}\}$

The three steeps are repeated $l = 1, \ldots, L$ times, where L denotes the number of Monte Carlo samples. The realizations $\{\pi_i^{(l)}(u \mid t) \, , \, l = 1, \ldots, L\}$ can be used to derive point estimates of $\pi_i(u \mid t)$, such as the median and the mean values:

$$\hat{\pi_i}^{(l)}(u \mid t) = \text{median}\{\pi_i^{(l)}(u \mid t) \, , \, l = 1, \ldots, L\} \tag{3.39}$$

$$\hat{\pi_i}^{(l)}(u \mid t) = L^{-1} \sum_{l=1}^{L} \pi_i^{(l)}(u \mid t) \tag{3.40}$$

From the estimates, it is also possible to compute the standard errors using the sample standard deviation over the Monte Carlo samples and the confidence intervals through the sample percentiles.

CHAPTER 4

A JOINT MODEL FOR THE PCA DATASET

4.1 Longitudinal analysis of PSA level over time

4.1.1 Specific random effects model

The general linear mixed model expression (Chapter 3, Section 3.2.3) can be specified as a particular mixed effects model, where each of the $n = 2415$ subjects from the `PCa Dataset` is modeled using subject-specific intercept and slope terms ($q = 2$). In our study, the observed longitudinal outcome corresponds to the *LLPSA* values for individual i at time t.

Since subjects are randomly sampled from a population, it is reasonable to assume that the subject-specific regression coefficients, $\tilde{\beta}_{i0}$ and $\tilde{\beta}_{i1}$, are also randomly sampled from the corresponding population of regression coefficients, with $(\tilde{\beta}_{i0}, \tilde{\beta}_{i1})^{\mathrm{T}} \sim \mathcal{N}_2((\beta_0, \beta_1)^{\mathrm{T}}, \mathbf{D})$, and the error terms are also assumed to be normally distributed, $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Therefore, the *LLPSA* subject profile is modeled as

$$LLPSA_i(t) = \tilde{\beta}_{i0} + \tilde{\beta}_{i1} t + \varepsilon_i(t) \ . \tag{4.1}$$

Each subject-specific coefficient can be separated into a fixed and random part, $\tilde{\beta}_{i0} = \beta_0 + b_{i0}$, $\tilde{\beta}_{i1} = \beta_1 + b_{i1}$, so our random effects model is finally expressed as:

$$\begin{cases} LLPSA_i(t) = \beta_0 + b_{i0} + (\beta_1 + b_{i1})t + \varepsilon_i(t) \\ \mathbf{b}_i \sim \mathcal{N}_2(\mathbf{0}, \mathbf{D}) \\ \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2) \end{cases} \tag{4.2}$$

where $LLPSA_i(t)$ is the outcome value for the $i-$th subject, $i = 1, \ldots, 2415$, at time t, $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$ is the fixed effects vector for the intercept and slope terms, $\mathbf{b}_i = (b_{i0}, b_{i1})^{\mathrm{T}}$ is the random effects vector for the intercept and slope, $\mathbf{D} = (d_{lk})_{l,k=1,2}$ is the 2×2 variance-covariance matrix, where $d_{11} = \mathrm{Var}(b_{i0}) = \sigma^2_{b_0}$, $d_{22} = \mathrm{Var}(b_{i1}) = \sigma^2_{b_1}$, and $d_{12} = \mathrm{Cov}(b_{i0}, b_{i1}) = \rho_{b_0 b_1} \sigma_{b_0} \sigma_{b_1}$, and $\varepsilon_i(t)$ is the sample error term on the $i-$th subject at time t.

We used an **unstructured covariance matrix** in the longitudinal analysis, that is, there are no constraints and each variance and each covariance is estimated uniquely from the data. This last assumption results in that variance and covariance values are very close to what the data reflect, and does not require knowledge or justification of a more restrictive pattern.

4.1.2 Analyzing response profiles

In order to determine any trends in the *LLPSA* mean response over time, a simple time plot can be obtained by connecting successive repeated measurements on the same subject with straight lines. However, it must be kept in mind that we do not know the subject's profile evolution between successive visits, as we only have information at specific time points.

Figure 4.1 shows the time-plots of *LLPSA*, where the consecutive observations on each man are connected through line segments. This plot offers a very good idea of the specific trajectories for all subjects, and it is colloquially known as *spaghetti plot* or *profile plot*. Interesting features on the referred plot are the wide variation in the initial values of *LLPSA* and the heterogeneity in the subsequent trajectories for different men. Also, it can be appreciated a marginal positive trend with a gentle slope, as well as a potential mildly nonlinear component.

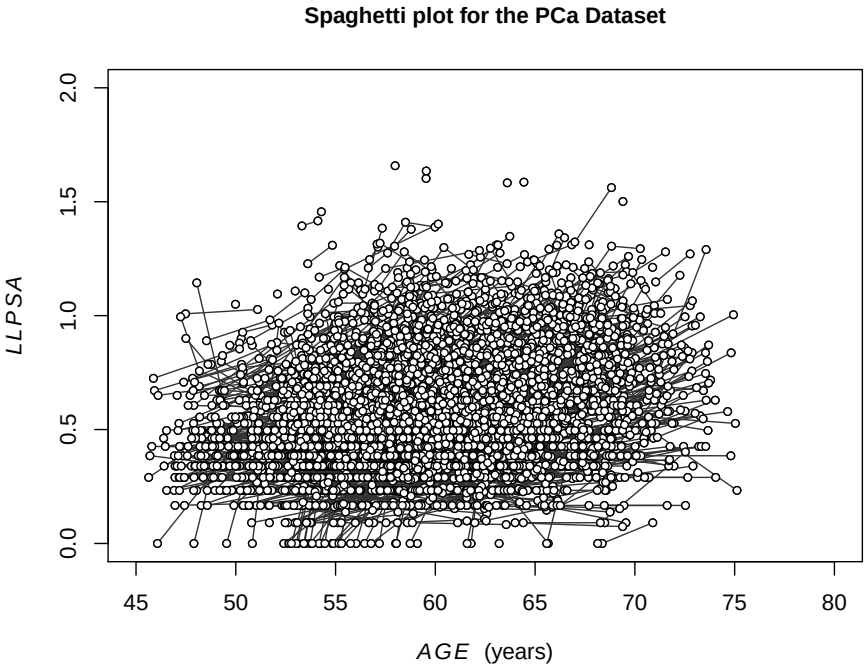

Figure 4.1. *LLPSA* subjects profiles across time (age in years) for the participants in the Spanish section of ERSPC.

However, in practice it is very difficult to track the response profile of any particular individual when the spaghetti plot contains so many individuals as in the above figure. As a result, it is more useful to present a time plot with joined line segments for only a relatively small and representative random sample selected from the study participants.

Figure 4.2 shows the *LLPSA* profile plots according to presence or absence of prostate cancer diagnosis. The top panel of the figure shows the *LLPSA* trajectories for the 116 men diagnosed with prostate cancer, while the bottom panel shows a random sample of 116 men chosen among the 2299 not diagnosed men. The first relevant aspect is that both panels indicated a similar behaviour in the general trend and in the variability of the *LLPSA* values. However, the prostate cancer group had higher mean values of *LLPSA*, in comparison with the non prostate cancer group. This information gives a clear support to the relevance of *PSA* measurements to explain the time to prostate cancer diagnosis.

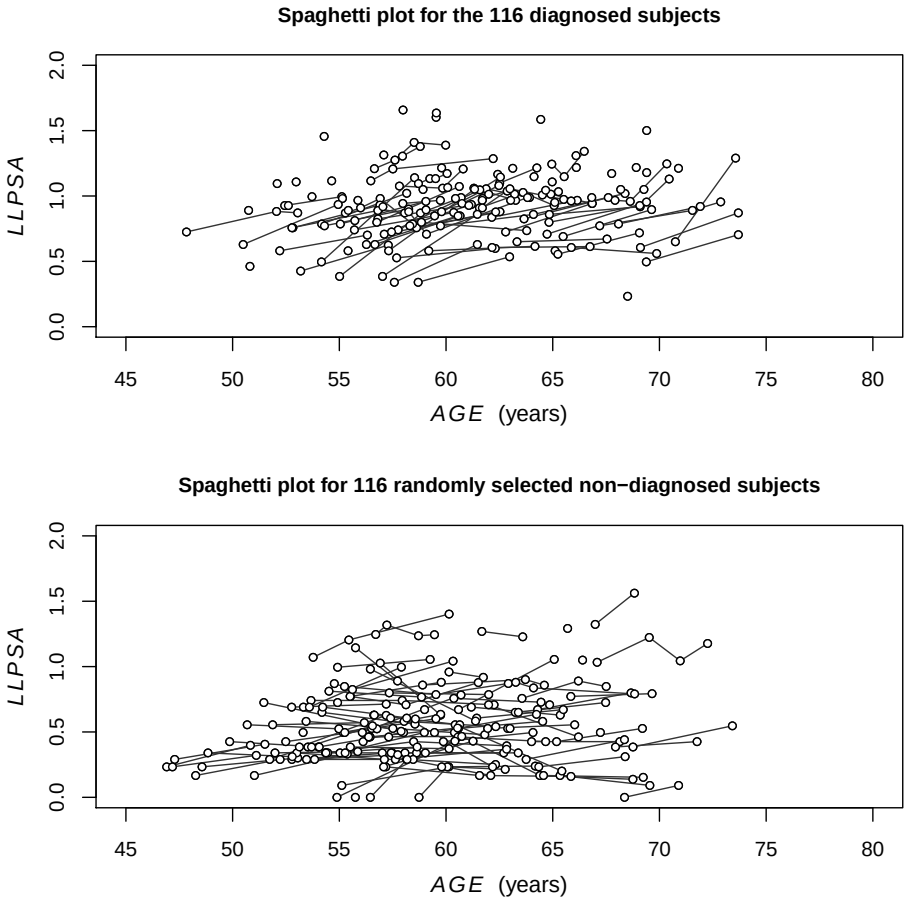

Figure 4.2. *LLPSA* subjects profiles across time (age in years) for the 116 men who developed prostate cancer during the follow-up period (top panel) and for 116 randomly selected men without a prostate cancer diagnosis (bottom panel), for the participants in the Spanish section of ERSPC.

4.1.3 Random effects model results for the PCa Dataset

Due to the high of variability of the subjects's responses at study entry, we first assumed random intercepts in the model,

$$LLPSA_i(t) = \beta_0 + b_{i0} + \beta_1 t + \varepsilon_i(t), \tag{4.3}$$

and then allow random intercepts and slopes,

$$LLPSA_i(t) = \beta_0 + b_{i0} + (\beta_1 + b_{i1})t + \varepsilon_i(t) \ . \tag{4.4}$$

In both cases, restricted maximum likelihood estimates are presented in Table 4.1.

The linear mixed model with random intercept assumes that each individual has a particular b_{i0}, and assumes a common slope. In this model there are two estimated variance components: $\hat{\sigma} = \sqrt{\widehat{\text{Var}}(\varepsilon_i(t))} = 0.107$ and $\hat{\sigma}_{b_0} = \sqrt{\widehat{\text{Var}}(b_{i0})} = 0.223$. The total variation (marginal variance) between any pair of *LLPSA* responses within the $i-$th subject is given by the sum of between-subject variability, $\hat{\sigma}^2_{b_0}$, and within-subject variability, $\hat{\sigma}^2$: $\widehat{\text{Var}}\{LLPSA_i(t)\} = \hat{\sigma}^2_{b_0} + \hat{\sigma}^2 = 0.223^2 +$

$0.107^2 = 0.247^2 = 0.061$, whereas the estimate marginal covariance between any pair of responses is $\widehat{\text{Cov}}\{LLPSA_i(t_j), LLPSA_i(t_k)\} = \hat{\sigma}^2_{b_0} = 0.223^2 = 0.050$.

As an example, we extract the estimated covariance matrix for subject $i = 556$ in the `PCa Dataset`, who has $n_{556} = 4$ visits at time points $\{t_{556,1} = 62.74,\ t_{556,2} = 66.93,\ t_{556,3} = 68.27,\ t_{556,4} = 69.58\}$ years,

Model	Estimate	Std. Error	95% Conf. Int.	p-value
Random intercept				
β_0	0.368	0.009	(0.349, 0.386)	< 0.0001
β_1	0.014	0.001	(0.013, 0.015)	< 0.0001
σ	0.107	–	(0.104, 0.110)	–
σ_{b_0}	0.223	–	(0.216, 0.231)	–
AIC	-2312.078	–	–	–
Random intercept and slope				
β_0	0.364	0.008	(0.347, 0.380)	< 0.0001
β_1	0.014	0.001	(0.013, 0.016)	< 0.0001
σ	0.105	–	(0.101, 0.108)	–
σ_{b_0}	0.147	–	(0.130, 0.166)	–
σ_{b_1}	0.006	–	(0.004, 0.008)	–
$\rho_{b_0 b_1}$	0.855	–	(-0.738, 0.998)	–
AIC	-2416.572	–	–	–

Table 4.1. Estimated parameters for the linear mixed effects models fit to the `PCa Dataset` by REML.

$$\hat{\text{V}}_{556,\text{ intercept}} = \begin{pmatrix} 0.061 & 0.050 & 0.050 & 0.050 \\ 0.050 & 0.061 & 0.050 & 0.050 \\ 0.050 & 0.050 & 0.061 & 0.050 \\ 0.050 & 0.050 & 0.050 & 0.061 \end{pmatrix}$$

Therefore, in the random intercept model, for subject $i = 556$ the percentage of total variation that is attributed to within-subject variability is $(0.107^2/0.247^2) \cdot 100 = 18.6\%$, with $(0.223^2/0.247^2) \cdot 100 = 81.4\%$ of total variation attributable to between-subjects variation in their general level of *LLPSA* (e.g., attributable to random intercepts).

The estimated correlation between any pair of measurements on the same subject is obtained by the named *intra-class correlation* $\widehat{\text{Corr}}\{LLPSA_i(t_j), LLPSA_i(t_k)\} = \hat{\sigma}^2_{b_0}/(\hat{\sigma}^2_{b_0} + \hat{\sigma}^2)$, so the estimated correlation matrix for subject $i = 556$ is:

$$\hat{\text{R}}_{556,\text{ intercept}} = \begin{pmatrix} 1 & 0.814 & 0.814 & 0.814 \\ 0.814 & 1 & 0.814 & 0.814 \\ 0.814 & 0.814 & 1 & 0.814 \\ 0.814 & 0.814 & 0.814 & 1 \end{pmatrix}$$

Additional flexibility is introduced in the random intercept and slope model, so that now the marginal variance of the $i-$th response is not constant but depends on the time point t, $\widehat{\text{Var}}\{LLPSA_i(t)\} = \hat{\sigma}^2_{b_0} + 2\widehat{\text{Cov}}(b_{i0}, b_{i1})t + \hat{\sigma}^2_{b_1}t^2 + \hat{\sigma}^2$, and also the marginal covariance between any pair of responses within the $i-$th subject, $\widehat{\text{Cov}}\{LLPSA_i(t_j), LLPSA_i(t_k)\} = \hat{\sigma}^2_{b_0} + 2\widehat{\text{Cov}}(b_{i0}, b_{i1})(t_{ij} + t_{ik}) + \hat{\sigma}^2_{b_0} t_{ij} t_{ik}$.

Again, the covariance matrix is obtained for the individual $i = 556$,

$$\hat{\mathrm{V}}_{556,\text{ intercept-slope}} = \begin{pmatrix} 0.069 & 0.064 & 0.065 & 0.067 \\ 0.064 & 0.081 & 0.072 & 0.074 \\ 0.065 & 0.072 & 0.085 & 0.076 \\ 0.067 & 0.074 & 0.076 & 0.089 \end{pmatrix}.$$

The total variation between any pair of *LLPSA* responses within the $i-$th subject is now $\hat{\sigma}^2_{b_0} + \hat{\sigma}^2_{b_1} + \hat{\sigma}^2 = 0.147^2 + 0.006^2 + 0.105^2 = 0.181^2 = 0.033$, and the between-subject variation accounts for $\{(\hat{\sigma}^2_{b_0} + \hat{\sigma}^2_{b_1})/(\hat{\sigma}^2_{b_0} + \hat{\sigma}^2_{b_1} + \hat{\sigma}^2)\} \cdot 100 = 66,4\%$ of all variability. From this result, a 99.8% is explained by the random intercept. It must be noted how the value of $\hat{\mathrm{Var}}(b_{i0})$ decreased in this model in comparison to the random intercept model, as now the model also relies on the random slope effect to explain between-subject variability.

The corresponding correlation matrix is derived from

$$\widehat{\mathrm{Corr}}\{LLPSA_i(t_j), LLPSA_i(t_k)\} = \frac{\hat{\sigma}^2_{b_0} + 2\widehat{\mathrm{Cov}}(b_{i0}, b_{i1})(t_{ij} + t_{ik}) + \hat{\sigma}^2_{b_0} t_{ij} t_{ik}}{\sqrt{\hat{\sigma}^2_{b_0} + 2\widehat{\mathrm{Cov}}(b_{i0}, b_{i1})t + \hat{\sigma}^2_{b_1} t_j^2 + \hat{\sigma}^2}\sqrt{\hat{\sigma}^2_{b_0} + 2\widehat{\mathrm{Cov}}(b_{i0}, b_{i1})t + \hat{\sigma}^2_{b_1} t_k^2 + \hat{\sigma}^2}}$$

$$\hat{\mathrm{R}}_{556,\text{ intercept-slope}} = \begin{pmatrix} 1 & 0.852 & 0.855 & 0.858 \\ 0.852 & 1 & 0.868 & 0.870 \\ 0.855 & 0.868 & 1 & 0.874 \\ 0.858 & 0.870 & 0.874 & 1 \end{pmatrix}$$

As expected, the general trend is that variances increase over time, whereas correlation decrease. It is evident that the random intercept and slope model offers greater flexibility in modeling the marginal covariance matrix, although it imposes a particular relationship that not necessarily is correct. Furthermore, it also draws attention the high correlation between the random effects, one may think that the random intercept effect is sufficient to explain the subject specific trend. However, $R^2_{b_0 b_1} = 0.731 < 0.80$, and therefore the correlation degree is not yet extremely high.

This model flexibility is also translated into an improvement that can be measured by the Akaike's Information Criterion, *AIC*. In this regard, a formal comparison is possible since the random intercept and slope model is nested into the random intercept model. The maximized *AIC* value was -2312.078 in random intercept model, while it decreases to -2416.572 when also allowing random slope. This decrease in *AIC* is quite substantial and statistically significant (ANOVA test) with p-value < 0.0001. So, in what follows in this section the longitudinal analysis is implemented with random intercept and slope model.

4.1.4 Subject specific predictions

A typical objective of longitudinal analysis is to characterize individual behavior. As mentioned above, the linear mixed effects model (which contains the random coefficient model as a special case) is a subject-specific model in the sense that an individual's "regression model" can be characterized at time t as having "mean" $\mathbf{x}_i^{\mathrm{T}}(t)\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}(t)\mathbf{b}_i$. Thus, to characterize the $i-$th specific subject profile with the random intercept and slope model, it is necessary to predict $\mathbf{b}_i = (b_{i0}, b_{i1})^{\mathrm{T}}$ for this subject in the model

$$\begin{cases} LLPSA_i(t) = 0.364 + b_{i0} + (0.014 + b_{i1})t + \varepsilon_i(t) \\ \mathbf{b}_i \sim \mathcal{N}_2(\mathbf{0}, \hat{\mathbf{D}}) \\ \varepsilon_i(t) \sim \mathcal{N}(0, \hat{\sigma}^2) \end{cases}$$

where $\hat{\mathbf{D}} = \begin{bmatrix} 0.147^2 & 0.001 \\ 0.001 & 0.006^2 \end{bmatrix}$ and $\hat{\sigma}^2 = 0.105^2$.

As already commented, the prediction of the random effects is carried out using the BLUP predictor.

To illustrate how the fitted mixed model works, five subjects with three visits (i.e., with sufficient longitudinal information) were randomly selected from our source dataset, $i = 100,\ 116,\ 128,\ 291,\ 543$. For each of them, a subject-specific profile was fitted by the predicted random effects, and the results were printed along with the global average response. The results are displayed in Figure 4.3 and show that the random effects model seems to fit adequately these individuals..

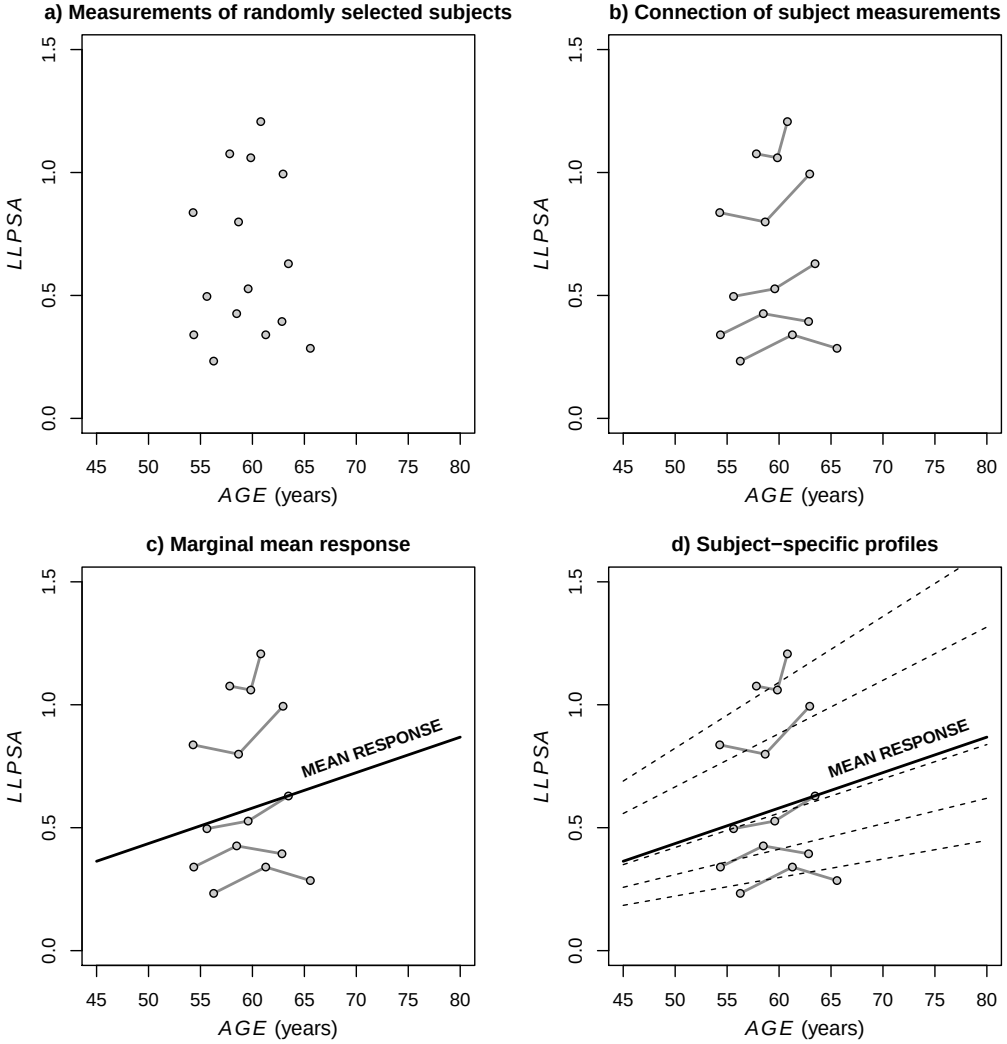

Figure 4.3. Representation of five subject specific profiles from the random intercept and slope model: a) *LLPSA* responses over time of the five selected subjects, b) Correlation between longitudinal responses, c) Average response across the individuals in the population, and d) Predicted subject-specific profiles.

A critical issue of the *PCa Dataset* is the existence of 573 subjects (23.7% of the total) with only one *LLPSA* measurement. For these special cases there is evidence that the model captures the information provided by subjects with more than one visit, assigning particular linear profiles whose slopes have a variability range with $\hat{\text{Var}}(b_1) = \sigma^2_{b1} = 0.006^2$. Figure 4.4 presents the predicted profiles for those 573 subjects without longitudinal follow-up are presented.

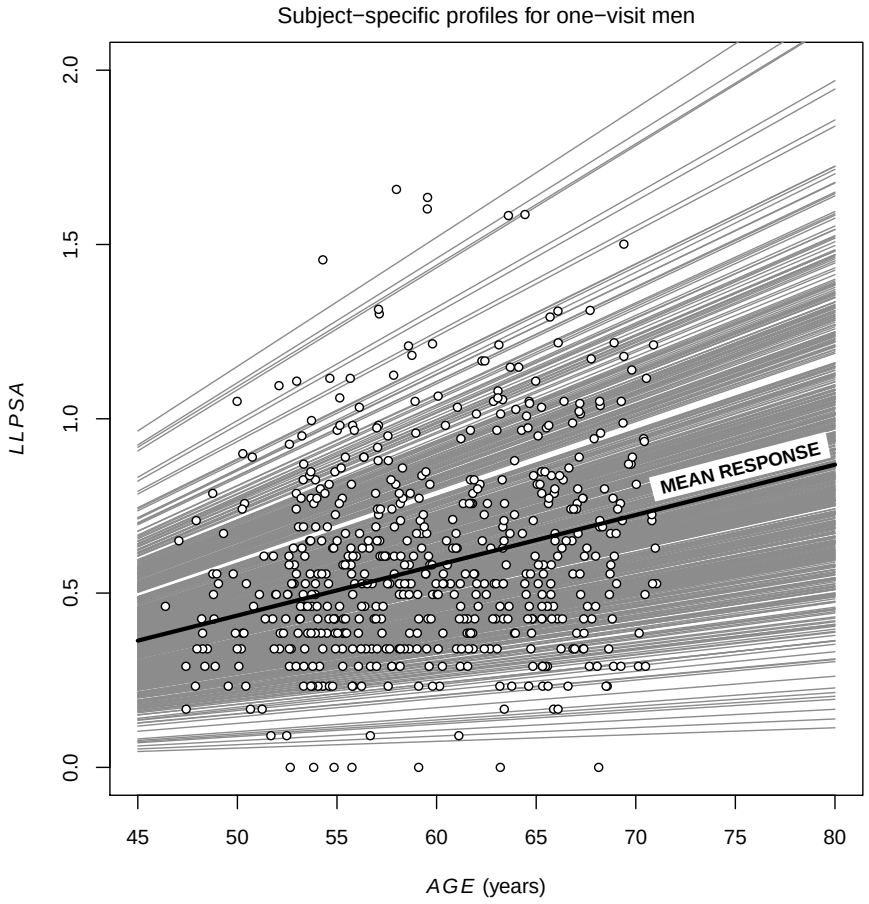

Figure 4.4. Prediction of subject-specific trens for the 573 individuals with only one *LLPSA* measurement.

Figure 4.4 shows that the slope allocation is not at random: the higher the *LLPSA* value, the greater the profile slope. The estimation is obtained using the information collected from the profiles with large number of visits.

4.2 Survival analysis of time to prostate cancer diagnosis

4.2.1 Survival results from non-parametric analysis

The Kaplan-Meier estimate is the simplest way of estimating survival over time. The Kaplan-Meier survival curve is defined as the probability of surviving a given period of time while considering time in many small intervals. There are three assumptions used in this analysis: 1) It is assumed that at any time subjects who are right-censored have the same survival prospects as those who continue to be followed, 2) The survival probabilities are the same for subjects recruited early and late in the study and 3) The event happens at the registered time.

Panel *a)* in Figure 4.5 shows the Kaplan-Meier (KM) estimate of the survival function of time to diagnosis for the study sample. The probability of prostate cancer diagnosis free survival decreases at a continuous rate of change in men aged 55 to 75 years and, due to censoring, it stabilizes at the estimated survival value 0.914. As previously mentioned, a $PSA \geq 3$ ng/ml is a key threshold in the screening program. Panel *b)* from the same plot displays the KM estimate of the survival function of time to PCa diagnosis for individuals whose first *PSA* measurement is below or above that limit. A rapid decrease of the survival function for individuals with the first $PSA \geq 3$ ng/ml is observed and a significant difference exists between the two categories (log-rank test with p-value < 0.0001). The estimated probability of prostate cancer diagnosis free survival for 75 years old subjects if their baseline *PSA* is below or above 3 ng/ml is 0.946 or 0.704, respectively.

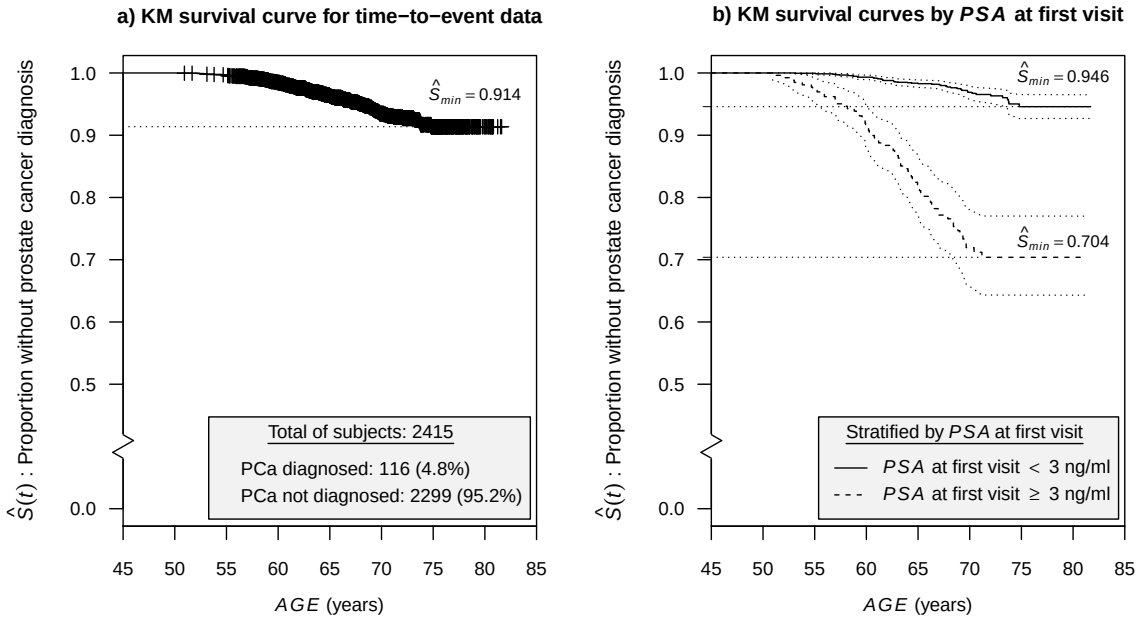

Figure 4.5. Plot of the Kaplan-Meier estimate of the survival function of time to prostate cancer diagnosis in the Spanish ERSPC study a) overall subjects in the sample, and b) stratified (and 95% confidence intervals) by the value (below or above 3 ng/ml) of the first *PSA* measurement.

Since time from study entry to prostate cancer diagnosis can vary depending on the *AGE*, it is important to account for this issue. Figure 4.6a displays KM survival curves of time from study entry to prostate cancer diagnosis, stratified by tertiles of age at entry ($[45, 55), [55, 60)$ and $\geq$60 years). An overall significant difference between the three categories is observed (log-rank p-value < 0.0001). Specifically, younger participants show better PCa diagnosis free survival estimates. However, since age and PSA levels are positively correlated, a joint analysis is needed. From this perspective, Figure 4.6b illustrates a significant difference between groups by $AGE \times PSA$ level at study entry (log-rank p-value < 0.0001), as well as a trend across the categories. Therefore, it will be necessary to consider the role of the interaction between *AGE* and *PSA* measurement at the first visit in the survival approach as a covariate to explain the time to event diagnosis.

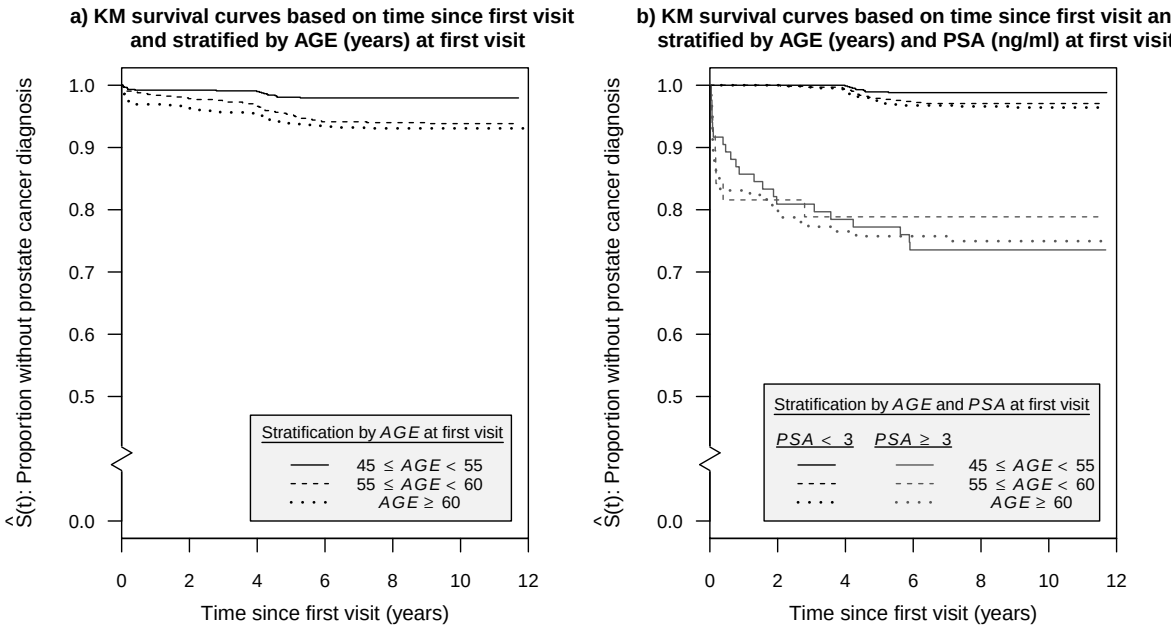

Figure 4.6. Plot of the Kaplan-Meier estimate of the survival function of time since the entry time to prostate cancer diagnosis in the Spanish ERSPC study a) stratified by age at entry, and (b) stratified by age and *PSA* level at entry.

4.2.2 Survival results for the Cox model

PH Cox model approach

As we have noted in the precedent subsection, there is an important point to explain the prostate cancer diagnosis in the degree of association between the $LLPSA$ level at first visit and the corresponding AGE, hereinafter called $LLPSA_0$ and AGE_0, respectively. These covariates have been treated as "baseline values". Taking this into account, for the $i-$th subject the following Cox PH model was fitted

$$h_i(t|\mathbf{w}_i) = h_0(t)R_i(t)\exp\{\gamma AGE_{0i} \times LLPSA_{0i}\}\ , \tag{4.5}$$

and the results displayed in Table 4.2 were obtained,

$\hat{\gamma}$	SE($\hat{\gamma}$)	HR	95% CI	p-value
0.119	0.012	1.126	(1.099, 1.153)	< 0.0001

Table 4.2. Estimated parameter of Cox PH model, $\hat{\gamma}$ with $AGE_{0i} \times LLPSA_{0i}$ as baseline covariate from `PCa Dataset`.

The fitted Cox PH model is

$$h_i(t|\mathbf{w}_i) = h_0(t)R_i(t)\exp\{\gamma AGE_{0i} \times LLPSA_{0i}\} = h_0(t)R_i(t)\exp\{0.119 AGE_{0i} \times LLPSA_{0i}\}. \tag{4.6}$$

The fact that the baseline covariate $AGE_0 \times LLPSA_0$ is significant and positive ($\hat{\gamma} = 0.119$, p-value $<$ 0.0001) implies that the $i-$th subject $LLPSA_{0i}$ level is significantly related with his corresponding

age, AGE_{0i}, and the risk increases with the $AGE_0 \times LLPSA_0$ values. Consequently, the hazard associated to a one-unit change of $LLPSA_{0i}$ not only depends of the recorded variation, but also of the baseline age at which the referred change takes place. As an illustration of the risk evolution, we can compare the hazard risk between two subjects that have the same $LLPSA_0$ level variation but at different baseline ages.

Let be one subject i who is 50 years ($AGE_{0i} = 5$) and increases his PSA_{0i} from 5 to 10 ng/ml (i.e., $LLPSA_0$ from 1.03 to 1.22),

$$\widehat{HR}_i = \frac{h_0(t)R_i(t)\exp\{0.119 \cdot 5 \cdot 1.22\}}{h_0(t)R_i(t)\exp\{0.119 \cdot 5 \cdot 1.03\}} = \exp\{0.119 \cdot 5 \cdot (1.22 - 1.03)\} = 1.12,$$

and other subject i' who experiences the same increase but at 65 years old ($AGE_{0i'} = 15$):

$$\widehat{HR}_{i'} = \frac{h_0(t)R_{i'}(t)\exp\{0.119 \cdot 15 \cdot 1.22\}}{h_0(t)R_{i'}(t)\exp\{0.119 \cdot 15 \cdot 1.03\}} = \exp\{0.119 \cdot 15 \cdot (1.22 - 1.03)\} = 1.40.$$

In this approach, if both subjects increase their corresponding prostate cancer diagnosis risk (as already known since $\hat{\gamma} > 0$), the older has a risk increase a $\{(1.40 - 1.12)/1.12\} \cdot 100 = 25\%$ higher than the younger one.

Extended Cox model approach

In practice, we can distinguish between the two types variables using the concept of predictability: an external or predictable variable is one whose value at any time t is known infinitesimally before t, whereas a endogenous or internal is not predictable. Standard time-to-event regression models (e.g., PH Cox model) treats time-dependent covariates like external, but this assumption is not true in the case of variables under biological processes as in the case of clinical biomarkers (e.g., *PSA*). The extended Cox Model works with the observed covariant history $\mathcal{Y}$, from which we only know the observed response at some specific points, and additionally these measures are associated to an measurement error. Thus, if we treat internal covariates as external, we may obtain biased results.

When handling with endogenous covariates, the extended Cox Model must be understood as an intermediate step towards a model that takes into account, at specific time t, the fully specification of the true variable path up to t, denoted by $\mathcal{M}(t)$ (Chapter 3, Subsection 3.4.1).

4.3 Joint modeling results

4.3.1 Estimation of joint model

We proceed by specifying and fitting the joint model that explicitly accounts for the endogeneity of the *LLPSA* marker. In particular, it has been taken into consideration two possible joint models depending on the longitudinal submodel adopted:

M1: A linear mixed model with random intercept for the longitudinal submodel and a relative risk

model including the baseline interaction between AGE and $LLPSA$, that is

$$\begin{cases} y_i(t) = m_i(t) + \varepsilon_i(t) = \beta_0 + b_{i0} + \beta_1 t \\ h_i(t|\mathcal{M}_i(t), \mathbf{w}_i) = h_0(t) R_i(t) \exp\{\gamma AGE_{01} \times LLPSA_{0i} + \alpha m_i(t)\} \end{cases} \tag{4.7}$$

where $b_{i0} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{b_0})$ and $\varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$.

M2: A linear mixed model with random intercept and slope for the longitudinal submodel and a relative risk model including the baseline interaction between AGE and $LLPSA$,

$$\begin{cases} y_i(t) = m_i(t) + \varepsilon_i(t) = \beta_0 + b_{i0} + (\beta_1 + b_{i1}) t \\ h_i(t|\mathcal{M}_i(t), \mathbf{w}_i) = h_0(t) R_i(t) \exp\{\gamma AGE_{01} \times LLPSA_{0i} + \alpha m_i(t)\} \end{cases} \tag{4.8}$$

where $\mathbf{b}_i \sim \mathcal{N}_2(\mathbf{0}, \mathbf{D})$, being $\mathbf{D}$ a unstructured 2×2 matrix for random effects, and $\varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$.

The joint model with longitudinal submodel M1 and longitudinal submodel M2 (right columns) parameter estimates are shown in Table 4.3 with their 95% respective confidence intervals. Inference focuses on the probability of the data (random) given the hypothesis (fixed) and the asymptotic properties of a specific estimator, under repeated sampling. As a result, confidence intervals for parameters can be derived.

	M1: Random intercept		M2: Random intercept and slope	
Parameters	Estimate	95% CI	Estimate	95% CI
Longitudinal submodel				
β_0	0.363	(0.345, 0.382)	0.362	(0.355, 0.369)
β_1	0.014	(0.013, 0.015)	0.014	(0.014, 0.015)
σ	0.107	–	0.082	–
σ_{b_0}	0.224	–	0.138	–
σ_{b_1}	–	–	0.005	–
$\rho_{b_0 b_1}$	–	–	0.999	–
Survival submodel				
γ	-0.048	(-0.086, -0.009)	-0.002	(-0.033, 0.030)
Association				
α	7.139	(6.001, 8.272)	5.490	(4.767, 6.213)
Goodness of fit				
AIC	-977.109	–	-8453.005	–

Table 4.3. Joint model estimates for ML analyses of longitudinal $LLPSA$ values and prostate cancer diagnosis. Two options have been considered for the longitudinal submodel: mixed model with random intercept and mixed model with random intercept and slope. The survival submodel follows a piecewise-constant model, and considers the $AGE_0 \times LLPSA_0$ interaction.

The parameter estimates for joint model (4.7) under the general unstructured random effects covariance matrix, showed a high correlation between the random effects, $\rho_{b_0 b_1} = 0.999$. To avoid unstability and colinearity in the estimates, we discarded the model that assumed non-independent random effects. Moreover, the joint model that assumed independent random effects did not provide a statistically significant $AGE_0 \times LLPSA_0$ interaction. Then, **we decided to include only**

the random intercept effect in the longitudinal approach and select the joint model with M1 as longitudinal submodel, with a relative risk model:

$$\begin{aligned} h_i(t|\mathcal{M}_i(t), \mathbf{w}_i) &= h_0(t)R_i(t)\exp\{\gamma AGE_{01} \times LLPSA_{0i} + \alpha m_i(t)\} = \\ &= h_0(t)R_i(t)\exp\{\gamma AGE_{0i} \times LLPSA_{0i} + \alpha(\beta_0 + b_{i0} + \beta_1 t)\}, \end{aligned} \tag{4.9}$$

where $b_{i0} \sim \mathcal{N}(0, \sigma^2_{b_0})$, and $\varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$.

With respect to the selected joint model (4.7), in contrast to the survival submodel result the sign for the baseline survival covariate is negative, $\hat{\gamma} = -0.048$, $p - value = 0.015$, so after combining longitudinal and survival processes, the prostate cancer risk decreases with the $AGE_0 \times LLPSA_0$ covariate. There is an increasing trend of prostate cancer risk as $AGE_0 \times LLPSA_0$ at study entry increases, which results from moderate or high values of the true profiles $m(t)$ slightly counterbalanced by the protective effect of the interaction.

Another important point from the results consists of noting both the positive value and the high significance of the association parameter, $\hat{\alpha} = 7.139$, $p - value < 0.0001$, so the *PSA* (expressed by *LLPSA*) level has a strong positive association with the risk for prostate cancer.

Table 4.4 presents the estimated HRs that correspond to different increases in *PSA* for a sample of *PSA* values. For instance, a unit increase in the *PSA* values represents a 4.7, 1.9 or 1.2-fold increase in the hazard rate of prostate cancer, for *PSA* values 1, 3 or 10, respectively. Therefore, the same changes in *PSA* have more impact on prostate cancer risk for low *PSA* values. However, for particular percentual increases of *PSA*, the HRs did not vary much across different *PSA* values. Increases of 10% or 20% in *PSA* levels result in HRs around 1.2 or 1.4, respectively.

		PSA level (ng/ml)						
		1	2	3	5	10	15	20
HR for an	1 unit	4.63	2.50	1.89	1.47	1.20	1.12	1.09
increase of	10%	1.22	1.24	1.24	1.22	1.20	1.18	1.17
PSA of	20%	1.48	1.51	1.50	1.47	1.41	1.37	1.35

Table 4.4. Estimated hazard ratio (HR) of prostate cancer for different increases of *PSA* at specific PSA values, under the joint model.

It is also important to point out that the joint model allows to assess the impact, by means of the hazard ratio $\exp[\hat{\alpha}\{\log(1 + PSA + \Delta PSA)) - \log(1 + \log(1 + PSA))\}]$ for a ΔPSA variation in *PSA* values depending on the *PSA* reference level. Figure 4.7 illustrates a contour plot of the estimated HR for a ΔPSA variation as a function of *PSA*. For instance, we can see that an increase of 2.0 ng/ml in *PSA* would represent a higher impact for a subject with 1.0 ng/ml of *PSA* ($\widehat{HR} = 10$) than for a subject whose *PSA* value was 8.0 ng/ml ($\widehat{HR} = 1.5$).

Next sections of this chapter focus on the joint model formulation for the `PCa Dataset`: A longitudinal submodel with random intercept and a survival submodel considering the interaction between AGE and LLPSA at study entry. From this model, the residuals validation and survival prediction results are obtained.

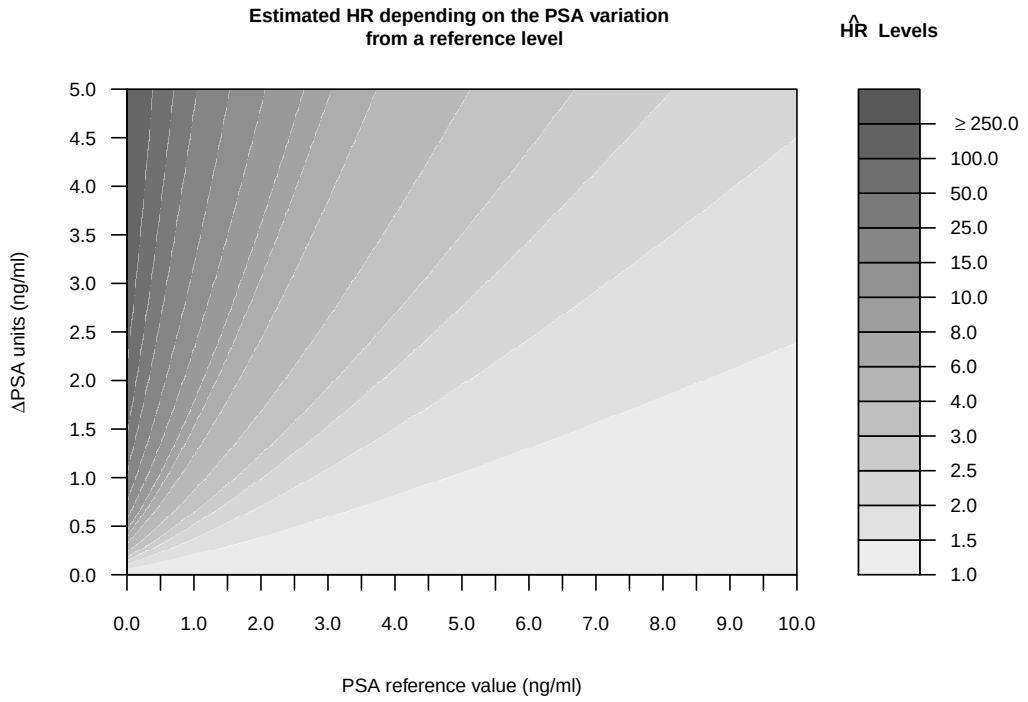

Figure 4.7. Contour plot of the estimated HR for a ΔPSA variation as a function of PSA.

4.3.2 Validation of the joint model

When it comes to using the joint model fitted for the `PCa Dataset`, a prerequisite step is to validate the model's assumptions. The standard and most frequently used tools to assess these assumptions are based on residual plots.

Diagnostic plots for the fitted joint model are shown in Figure 4.8. The two top charts provide the residual plots to validate the longitudinal submodel, whereas the bottom plots are used to assess the adequacy of the survival submodel.

Concerning to the standard linear mixed-effects model (i.e., the longitudinal part), the subject-specific residuals plot (top left panel) predicts the conditional errors at the specified time points, $\varepsilon_i(t_{ij})$, which seem to fulfill the homoscedasticity and normality assumptions. In addition, the fitted loess curve in the standardized marginal residuals plot (top right panel) displays no systematic pattern, with a random scatter around a constant zero mean, which confirms the suitability of the structure for the fixed effects design $(n_i \times p)$-dimensional matrix for a given subject, $\mathbf{X}_i$.

The survival marginal residuals (bottom left panel) show a slight deviation of the loess smother from zero and confirm the appropriateness of the chosen functional form for the PSA values. The bottom right panel, that shows the comparison between the KM curve for the Cox-Snell residuals and the unit exponential distribution, detects only some lack of fit for residual values on the right tail of the distribution.

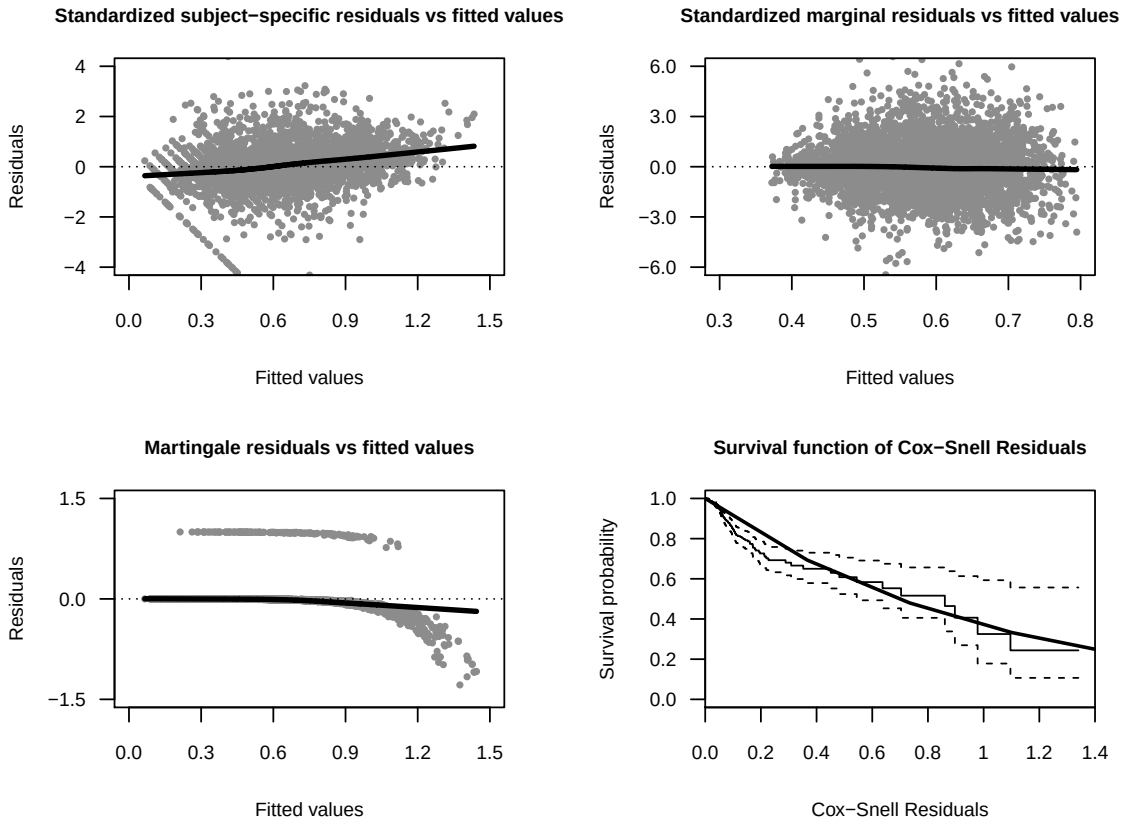

Figure 4.8. Diagnostic plots for the longitudinal submodel (two top panels) and the survival submodel (two bottom panels) for the fitted joint model.

4.3.3 Dynamic predictions of survival probabilities

In this section we focus on expected survival for all the 2415 subjects from the `PCa Dataset`, but within the joint modelling framework. In particular, based on the joint model fitted, it would be possible to predict survival probabilities for a new subject that has provided a set of longitudinal measurements.

As already explained, the fitted joint model allows to obtain individual dynamic predictions of prostate cancer free survival probabilities based on the observed longitudinal profile at specific time points. As an example to illustrate how changes in the *LLPSA* profiles are reflected in changes in the dynamic updates of the survival probabilities, Figure 4.9 shows how the estimated survival curve for the subject $i = 556$ from `PCa Dataset` is updated as the number of *PSA* measurements grows from one to four at the corresponding visits: $\{t_{556,1} = 62.74,\ t_{556,2} = 66.93,\ t_{556,3} = 68.27,\ t_{556,4} = 69.58\}$ years. The idea behind this four-panel graph consists of including at each time point the last available *LLPSA* measurement and the estimated survival curve after this point.

In our case, the median estimator was employed to represent survival probabilities. From the four plots included in the figure, there is a constant increase in the true level of *LLPSA*, leading to a lower prostate cancer free survival probability, which highlights the relevance of an accurate follow-up when aiming implement an individualized screening approach.

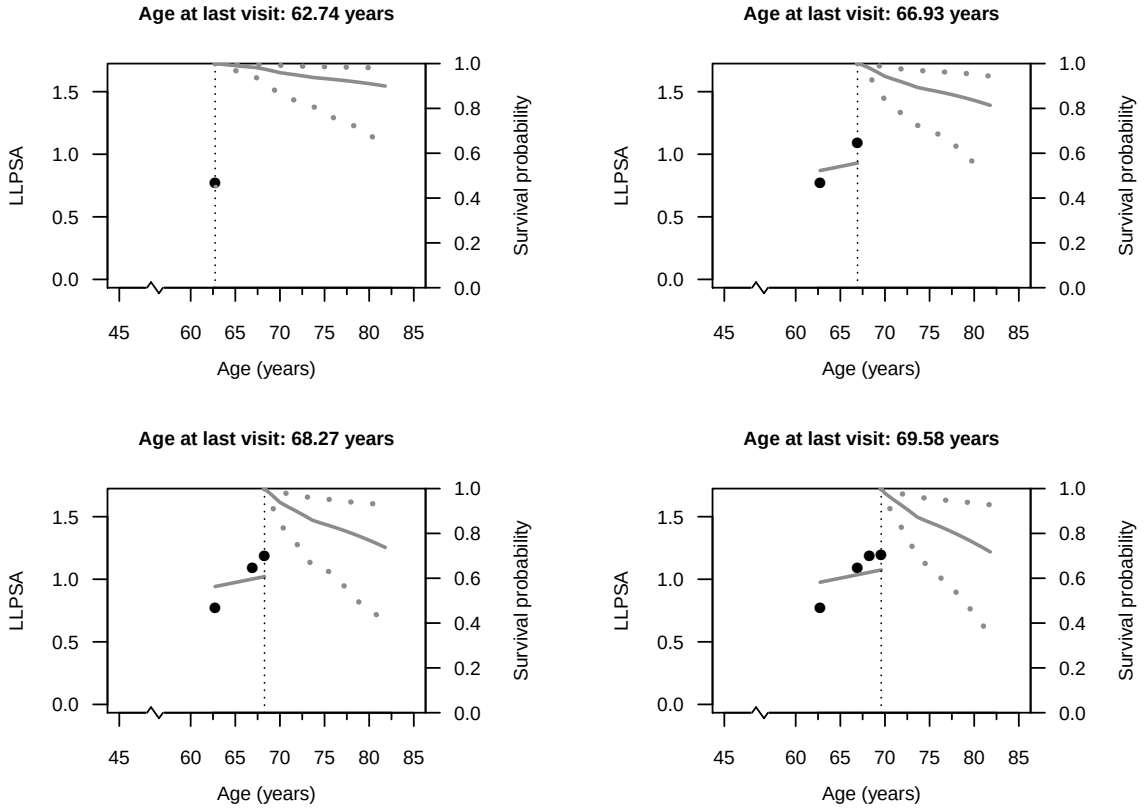

Figure 4.9. Successive LLPSA longitudinal trajectories and dynamic prostate cancer free survival probabilities (median estimator and 95% pointwise confidence intervals for $\pi_i(u|t)$ at each time point), for subject $i = 556$ from the `PCa Dataset`.

Therefore, it is possble to obtain the survival probabilities for subject $i = 556$, of whom we know that he was free diagnosed until the end of the study (when he was 81.80 years old), although his last *PSA* measurement took place at age of 69.58 (4.10).

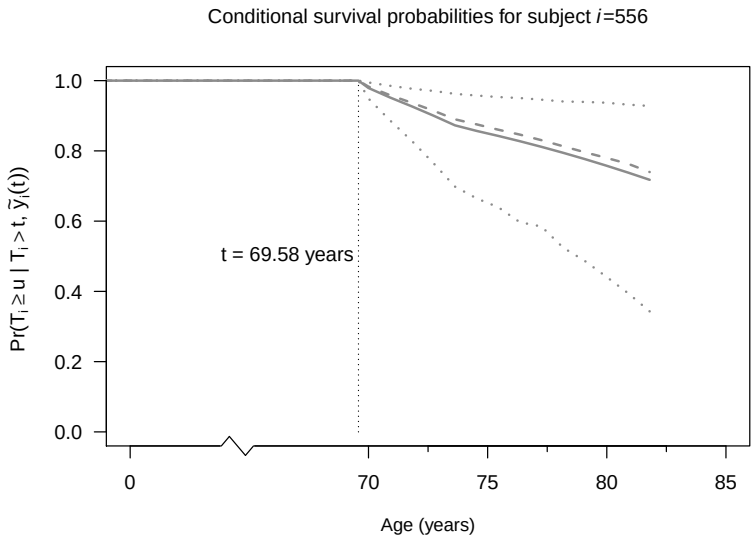

Figure 4.10. Survival probabilities for subject $i = 556$ from the `PCa dataset`. The solid and dashed lines correspond to the median and mean estimators, respectively.

CHAPTER 5
DISCUSSION AND FUTURE RESEARCH

5.1 Discussion and Conclusions

In recent years, clinical researchers have shown a great interest to record the values of key longitudinal covariates until the occurrence of a particular event in a subject. However, such covariates are usually associated to inherent biological processess of each subject, so that a separate analyses of longitudinal and survival data may lead to inefficient or biased results. In order to overcome these protential difficulties, it is necessary to advance towars a jointly modelization of both approaches. In this line, our study main goal has been to understand and apply the necessary statistical concepts to model adequatly endogenous time-dependent covariates. In this regard, we have illustrated the capabilities of the joint modelling for longitudinal and time-to-event data using shared parameter models. These models are applicable either to account for the effect of a time-dependent covariate measured with error in a survival analysis context or to correct for non at random dropout in the analysis of longitudinal outcomes. To choose methods for inference, the joint likelihood method generally produces most reliable results if the assumed models and distributions are correct. In particular, a joint model has been fitted successfully to the motivating dataset of the study, consisting in 2415 monitoring subjects from the Spanish branch of ERSPC study. The main research question was to explain the prostate cancer diagnosis risk on each of these subjects as a result of the repeated measurements over time of the *PSA* biomarker, considering positive correlation between the responses on the same subject.

The results showed that our fitted model is consistent with the literature and clinical knowledge. On the one hand, the observed *PSA* is highly associated with the risk of being diagnosed of PCa. On the other hand, there is a protective effect of the age and *PSA* interaction, consistent with the age-varying effect of *PSA* on prostate cancer risk. Our study consisted of using joint modeling techniques to 1) obtain an unbiased and efficient estimate of the the impact of *PSA* trajectories on time to prostate cancer diagnosis, and 2) refine the prostate cancer free survival estimates by using dynamic predictions based on the whole true history of *PSA* evolution for each subject.

The fitted joint model was validated by residuals tecniques, and predicted survival probabilities were calculated. As longitudinal *PSA* information was collected for all subjects, the joint modeling methodology has allowed to continuously update the predictions of their survival probabilities, and therefore being able to discern between men with low and high risk for a disease diagnosis.

5.2 Future research

Our work can be considered a first step that should be followed by a more comprehensive modeling process accounting for additional predictive factors measured over time like digital rectal examination, ultrasound biopsy test, prostate volume, family history and previous biopsy status among others. In this respect, a very important aspect would be to have other baseline covariates in further studies on this issue.

Another important aspect from our data is that the number of observations per subject was small, so

that only 14% of them had three or more *PSA* measurements. More within subject measurements would have prevented some potential bias in the random effect estimate and also would have produced more precise estimates of the association parameter.

Even though we have focused on joint models with a relative risk submodel with a piecewise constant baseline risk function for the event outcome, `JM` package form `R` software offers several other options for the survival submodel as described in Section 3.4, such as Weibull or Gamma distributions.

Finally, it must be pointed out that extensions of the standard joint model approach can be considered. Thus, it is reasonably to consider not only the biomarker true value at specific time point t to predict the subject-specific hazard, but also the particular slope (trend) of the longitudinal trajectory up to this time.

Bibliography

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10:1100–1120.

Andriole, G. L., Crawford, E. D., Grubb, R. L., Buys, S. S., Chia, D., Church, T. R., Fouad, M. N., Gelmann, E. P., Kvale, P. A., Reding, D. J., Weissfeld, J. L., Yokochi, L. A., O'Brien, B., Clapp, J. D., Rathmell, J. M., Riley, T. L., Hayes, R. B., Kramer, B. S., Izmirlian, G., Miller, A. B., Pinsky, P. F., Prorok, P. C., Gohagan, J. K., Berg, C. D., and Team, P. P. (2009). Mortality results from a randomized prostate-cancer screening trial. *The New England Journal of Medicine*, 360:1310–1319.

Berenguer, A., Luján, M., Páez, A., Santonja, C., and Pascual, T. (2003). The Spanish contribution to the European Randomized Study of Screening for Prostate Cancer. *BJU international*, 92:33–38.

Brandt, L. J., Sheng, S. L., Morrell, C. H., Verbeke, G. N., Lesaffre, E., and Carter, H. B. (2003). Screening for prostate cancer by using random-effects models. *Journal of the Royal Statistical Society: Series A*, 166:51–62.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30:89–99.

Carter, H. B., Albertsen, P. C., Barry, M. J., Etzioni, R., Freedland, S. J., Greene, K. L., Holmberg, L., Kantoff, P., Konety, B. R., Murad, M. H., Penson, D. F., and Zietman, A. L. (2013). Early detection of prostate cancer: AUA Guideline. *American Urological Association*, 190:419–426.

Chou, R., Croswell, J. M., Dana, T., Bougatsos, C., Blazina, I., Fu, R., Gleitsmann, K., Koenig, H. C., Lam, C., Maltz, A., Rugge, J. B., and Lin, K. (2011). Screening for prostate cancer: a review of the evidence for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*, 155:762–771.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34:187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62:269–276.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72:557–565.

Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine*, 15:1663–1685.

Ferlay, J., Shin, H., Bray, F., Forman, D., Mathers, C., and Parkin, D. (2010). GLOBOCAN 2008, cancer incidence and mortality worldwide: IARC CancerBase No. 10. *Lyon, France: International Agency for Research on Cancer*, 2010:29.

Garre, F. G., Zwinderman, A. H., Geskus, R. B., and Sijpkens, Y. W. (2008). A joint latent class changepoint model to improve the prediction of time to graft failure. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171:299–308.

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69:553–566.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72:320–338.

Heijmink, S. W., Fütterer, J. J., Strum, S. S., Oyen, W. J., Frauscher, F., Witjes, J. A., and Barentsz, J. O. (2011). State-of-the-art uroradiologic imaging in the diagnosis of prostate cancer. *Acta Oncólogica*, 50:25–38.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1:465–480.

Hsieh, F., Tseng, Y.-K., and Wang, J.-L. (2006). Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics*, 62:1037–1043.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data, 2nd Edition*, volume 360. John Wiley & Sons.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:963–974.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83:1014–1022.

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Gelber, R., and Lipshultz, S. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics*, 58:621–630.

Little, R. J. and Rubin, D. B. (2002). *Statistical analysis with missing data, 2nd edition*. Wiley, New York.

Luján, M., Páez, A., Berenguer, A., and Rodriguez, J. A. (2012). Mortality due to prostate cancer in the Spanish arm of the European Randomized Study of Screening for Prostate Cancer (ERSPC). Results after a 15-year follow-up. *Actas Urológicas Españolas*, 36:403–409.

Moyer, V. A. (2012). Screening for prostate cancer: US Preventive Services Task Force recommendation statement. *Annals of Internal Medicine*, 157:120–134.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and Team, R. C. `nlme`: Linear and nonlinear mixed effects models, 2012. *R package version*, pages 3–1.

Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69:331–342.

Proust-Lima, C. and Taylor, J. M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of post-treatment PSA: A joint modeling approach. *Biostatistics*, 10:535–549.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35:1–33.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67:819–829.

Rizopoulos, D. (2012a). JMbayes: Shared parameter models for the joint modeling of longitudinal and time-to-event data using JAGS, WinBUGS, or OpenBUGS.

Rizopoulos, D. (2012b). *Joint models for longitudinal and time-to-event data with applications in R*. CRC Press, Boca Ratón, FL.

Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:637–654.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2010). Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics*, 66:20–29.

Roobol, M. J., van Vugt, H. A., Loeb, S., Zhu, X., Bul, M., Bangma, C. H., van Leenders, A. G., Steyerberg, E. W., and Schroder, F. H. (2012). Prediction of prostate cancer risk: the role of prostate volume and digital rectal examination in the ERSPC risk calculators. *European Urology*, 61:577–583.

Santos Nobre, J. and da Motta Singer, J. (2007). Residual analysis for linear mixed models. *Biometrical Journal*, 49:863–875.

Schabenberger, O. and Pierce, F. F. J. (2002). *Contemporary statistical models for the plant and soil sciences*. CRC press.

Schröder, F. H., Hugosson, J., Carlsson, S., Tammela, T., Määttänen, L., Auvinen, A., Kwiatkowski, M., Recker, F., and Roobol, M. J. (2012). Screening for prostate cancer decreases the risk of developing metastatic disease: findings from the European Randomized Study of Screening for Prostate Cancer (ERSPC). *European Urology*, 62:745–752.

Schröder, F. H., Hugosson, J., Roobol, M. J., Tammela, T. L., Ciatto, S., Nelen, V., Kwiatkowski, M., Lujan, M., Lilja, H., Zappa, M., Denis, L. J., Recker, F., Paez, A., Maattanen, L., Bangma, C. H., Aus, G., Carlsson, S., Villers, A., Rebillard, X., van der Kwast, T., Kujala, P. M., Blijenberg, B. G., Stenman, U. H., Huber, A., Taari, K., Hakama, M., Moss, S. M., de Koning, H. J., and Auvinen, A. (2012). Prostate-cancer mortality at 11 years of follow-up. *The New England Journal of Medicine*, 366:981–990.

Slate, E. H. and Turnbull, B. W. (2000). Statistical models for longitudinal biomarkers of disease onset. *Statistics in medicine*, 19:617–637.

Swerdlow, S., Campo, E., Harris, N., et al. (2008). International Agency for Research on Cancer. *WHO classification of tumours of haematopoietic and lymphoid tissue*, pages 1–439.

Therneau, T. (2012). Survival analysis, including penalised likelihood. R package.

Therneau, T. and Grambsch, P. (2000). *Modeling survival data: extending the Cox model*. Springer, New York.

Tsiatis, A., Degruttola, V., and Wulfsohn, M. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90:27–37.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, 14:809–834.

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53:330–339.

Appendix A
DETAILS ON SOURCE DATASET CONFIGURATION

A.1 The two types of observations in `PCa Dataset`

In order to collect all the information provided by the ERSPC Spanish section, some artificial observations (as additional rows) were introduced in our source data set. Thereby, there are two types of observations:

1. Observations that correspond to visit dates

   These correspond to observations that traditionally may occur in any database where there are multiple records per person. Such observations correspond to regular monitoring visits, with their corresponding *PSA* level and with known codes for the *DRE* and *TRUS* physical tests. These rows are easily identifiable in our dataset because they are always associated with two possible code pairs:

   ($RECTYP = 0$; $BIOPSY = 0$): Visit in which a biopsy test was not recommended.

   ($RECTYP = 0$; $BIOPSY = 1$): Visit in which biopsy test was recommended but not done.

   The total number of such rows in `PCa Dataset` is 4673, unequally distributed among the 2415 subjects involved in the study. **In practice, this is the information that, adapted in the most convenient way for each analysis, has been used throughout this study in order to describe correctly the survival and longitudinal approaches**.

2. Artificial observations

   These are dataset rows which were "artificially" added to a given subject with the only purpose of reporting the date of a specific event occurred in the historical profile of the subject. Possible reasons that may lead to the addition of an artificial observation within specific subject were the following:

   - Observation that reports the date of a performed prostate biopsy:
     - An observation which indicates the date that a subject underwent biopsy with negative result is coded as ($RECTYP = 0$; $BIOPSY = 2$).
     - An observation which indicates the diagnosis date of a screen-detected prostate cancer is coded as ($RECTYP = 0$; $BIOPSY = 3$).
     - An observation which indicates the diagnosis date of an interval prostate cancer is coded as ($RECTYP = 0$; $BIOPSY =$NA).
   - Observation that reports the date of the subject's exclusion date from trial or his death date, without having in both cases any relationship with prostate cancer event:
     - An observation which indicates the exclusion date from trial is coded by $RECTYP = 1$.
     - An observation which indicates the death date is coded by $RECTYP = 2$.

The total number of artificial rows in `PCa Dataset` is 1121, unequally distributed among the 2415 subjects involved in the study. **This kind of rows were not finally included in the joint modeling analysis carried out, although they are a very valuable data, both to better understand the patterns of disease and for offering the basis for further work on this issue**.

A.2 Treatment of exclusions from the study

As already mentioned, the observations that only report exclusion dates are purely informative, so the active active surveillance of prostate cancer incidence continued until 31/12/2007 (provided that the exclusion had occurred before this study end).

Under the above assumption, there were recorded a total of 249 exclusions prior to study end on 31/12/2007, which were properly incorporated to the `PCa Dataset` as "artificial" rows using the code $RECTYP = 1$.

On the other hand, during the data set debugging six individuals were found with an ordinary visit recorded after his exclusion date. Since exclusion involves the ending of re-screening visits, these observations were removed from our source data set alleging that that an exclusion is a more rational criterion than a visit. In this respect, it must also be pointed out that three of the removed visits were associated with not done biopsy recommendations, while the remaining three visits resulted respectively in three negative biopsy tests.

Finally, the number of recommended but not done biopsies became 278 in our `PCa Dataset` versus the 281 in the Spanish section from ERSPC trial, while negative biopsies were 584 versus the 587 listed in the mentioned trial.

A.3 Treatment of missing values

Some of the variables introduced in `PCa Dataset` contained missing values for certain observations, both in visit rows and artificially added rows. Throughout the entire data set, these missing values were coded using NA, but for the correct understanding of the `PCa Dataset` it is important to note the row type within which the missing value appears:

1. Missing data that corresponds to an unknown value in a row of a specific subject:

   These would be the missing values that traditionally may occur in any database. The reason why in some records from our database there are unknown values for some variables, is only due to available instruments or to protocol issues, so one can assume a pattern of the type *Missing Completely at Random* (MCAR), without any relationship to development of prostate cancer.

2. Missing data from record dates (rows) that have been "artificially" added to a given subject:

   These rows had to be included in order to record the exact date of some specific events described earlier. They are observations that do not represent an individual's visit, and subsequently they can not provide any information related to the covariates of interest. For this reason, it was decided to codify with NA these above mentioned covariates.

A.4 Biopsies results' distribution

The study involves active monitoring of the potential prostate cancer incidence in subjects through sequential readings of their *PSA* level. Thereby, the subject was recommended a biopsy if his *PSA* level was above the threshold established at the time of the visit. From the date of the first study visits, on 19/02/1996, until the last visits day, on 21/10/2005, a total of 968 biopsies were recommended over the course of 4673 occurred visits. Biopsies results' distribution is shown in Table A.1:

Biopsies performed		684 (71.1%)
Biopsies with negative result	584 (85.4% / 60.7%)	
Biopsies with positive result	100 (14.6% / 10.4%)	
Biopsies recommended but not done		278 (28.9%)
Total		**962 (100%)**

Table A.1. Distribution of the biopsies' results in the `PCa Dataset`.

The total number of biopsies performed in the study were 684: 584 (85.4%) from these had a negative result and there were 100 (14.6%) which led to a prostate cancer diagnosis. Moreover, there were 278 recommended but not done biopsies distributed among 242 different patients, so 10.0% of total 2415 subjects had at least one record with a recommended but not performed biopsy.

Finally, the distribution of 116 diagnosed prostate cancer cases was the following:

- Number of individuals with screen-detected prostate cancer: 100
- Number of patients with interval prostate cancer diagnosed: 16. From these group, there were 15 interval cancers diagnosed, whom joined one subject who did not present a positive biopsy within the study but had prostate cancer as the cause of death in the death certificate.

A.5 Examples of profile description

Operating scheme of data set can be complex, specially if the reader is not familiar with it. Consequently, this section proposes, by way of illustration, the description of the historical profile of three subjects contained in the database. For this, their path were covered from first visit date until its last record due to one of different possible ending scenarios.

The three subjects' profiles chosen for illustration are 100, 138 and 275, so that their respective associated records are shown in Table A.2:

ID	*RANDAT*	*DATE*	*RECTYP*	*AGE*	*PSA*	*DRE*	*TRUS*	*BIOPSY*	*FPSA*	*PSARAT*	*TIME*	*CENS*	*CANTYP*	*GLE2*
⋮														
100	14/05/1996	24/05/1996	0	61.31	2.70	0	0	0	NA	NA	72.92	0	0	0
100	14/05/1996	21/06/2001	0	66.39	5.10	0	0	1	NA	NA	72.92	0	0	0
100	14/05/1996	20/11/2002	0	67.81	6.22	1	1	0	11.20	1.80	72.92	0	0	0
100	14/05/1996	19/12/2002	0	67.89	NA	NA	NA	2	NA	NA	72.92	0	0	0
100	14/05/1996	25/02/2004	0	69.07	5.56	1	1	0	1.39	0.25	72.92	0	0	0
100	14/05/1996	09/03/2004	0	69.11	NA	NA	NA	2	NA	NA	72.92	0	0	0
⋮														
138	02/06/1996	12/06/1996	0	64.82	0.70	0	0	0	NA	NA	69.54	0	0	0
138	02/06/1996	17/12/1999	1	68.34	NA	NA	NA	NA	NA	NA	69.54	0	0	0
138	02/06/1996	02/03/2001	2	69.54	NA	NA	NA	NA	NA	NA	69.54	0	0	0
⋮														
275	30/10/1996	06/11/1996	0	58.62	2.10	0	0	0	NA	NA	62.62	1	1	1
275	30/10/1996	04/10/2000	0	62.53	3.10	1	1	0	NA	NA	62.62	1	1	1
275	30/10/1996	08/11/2000	0	62.62	NA	1	NA	3	NA	NA	62.62	1	1	1
⋮														

Table A.2. Records of individuals 100, 138 and 275 in the `PCa Dataset`.

Personal history profile of subject 100:

The individual 100, assigned to the screening group on 14/05/1996, had his first visit on 24/05/1996 at the age of 61.31 years, with $PSA = 2.70$ ng/ml registered. According to the first protocol, required neither biopsy (because his $PSA \leq 4$ ng/ml) nor an early recall ($PSA \notin [3, 4]$ ng/ml). He had a new visit on 21/06/2001 (later of theoretically four years between standard visits), and recorded a total PSA level greater than the threshold under the second protocol: $PSA = 5.10$ ng/ml ≥ 3 ng/ml. He was then recommended a biopsy test but he rejected.

Due to the high serum level recorded in his second visit (and to the fact that he did not have a biopsy), he was assigned to an early recall (within less than two years) which took place on 20/11/2002. He recorded then total PSA and $PSARAT$ values above the thresholds established in 3rd protocol: $PSA = 6.22$ ng/ml ≥ 3 ng/ml and $PSARAT = 1.80 \geq 0.20$. Thereby, he underwent a biopsy on 19/12/2002, obtaining a negative prostate cancer diagnosis.

Since the high PSA value on his third visit and the subsequent negative biopsy result, he was send to a new early visit on 25/02/2004, in which he was again recorded PSA and $PSARAT$ values higher than the maximum allowable thresholds: $PSA = 5.56$ ng/ml≥ 3 ng/ml and $PSARAT = 0.25 \geq 0.20$. Consequently, he was recommended a new biopsy test, which took place on 9/3/2004 and had a negative result in relation to prostate cancer. Afterwards, the individual continued his stay in the study until 31/12/2007, when he was 72.92.

After this history profile, it can be determined that the follow-up of the subject 100 extends from the date of his first visit, at the age of 61.31 years, to the study end, aged 72.92 years, not having been diagnosed with prostate cancer diagnosis during this period (this fact is represented with a circle at next scheme).

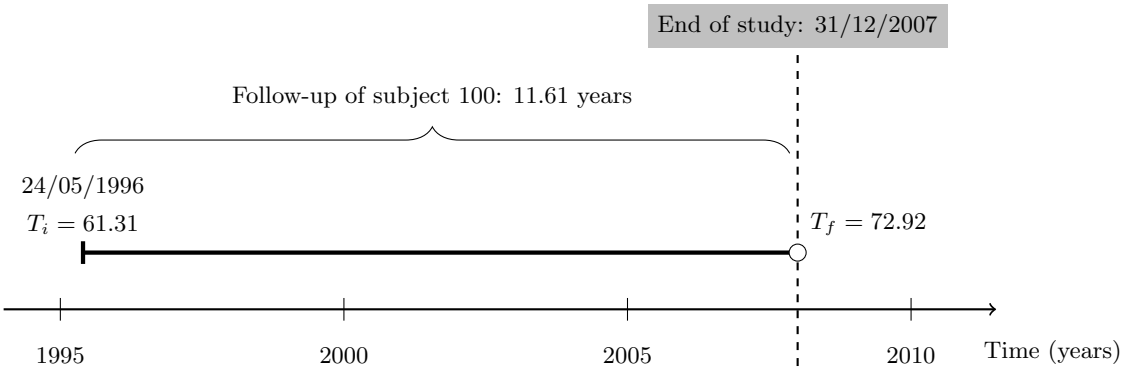

Figure A.1. Subject 100 folow-up within the observation window.

Personal history profile of subject 138:

The individual 138, assigned to the screening group on 02/06/1996, had his first and only visit on 12/06/1996 at the age of 64.82, with $PSA = 0.70$ ng/ml. According to the first protocol, he required neither biopsy (he had a $PSA \leq 4$ ng/ml) nor an early recall ($PSA \notin [3, 4]$ ng/ml).

On 17/12/1999, this subject was excluded from further screening rounds due to unrelated causes with prostate cancer, but continued under an active surveillance on the potential incidence of the disease.

Finally, the subject died on 02/03/2001 at the age of 69.54 years without any prostate cancer

symptoms.

After this information, it can be determined that the follow-up period of the subject 138 extends from the date of his first visit until the date of his death (prior to the end of study), with no signs of prostate cancer.

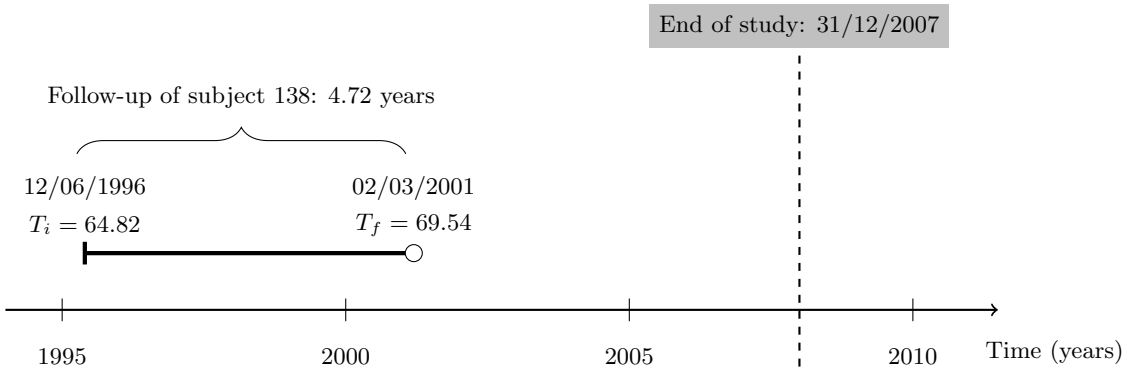

Figure A.2. Subject 138 folow-up within the observation window.

Personal history profile of subject 275:

The individual 275, assigned to the screening group on 30/10/1996, had a first visit on 06/11/1996 at the age of 58.62 years with $PSA = 2.10$ ng/ml. According to the first protocol required neither biopsy (he had a $PSA \leq 4$ ng/ml) or an early recall ($PSA \notin [3, 4]$ ng/ml).

He had a new visit on 04/10/2000 (four years later, as standard interval screenings indicates), and recorded a total PSA level greater than the threshold under the second protocol: $PSA = 3.10$ ng/ml $\geq$ 3ng/ml. He was then recommended a biopsy test, performed on 08/11/2000 and with a positive result: prostate cancer diagnosed when he was 62.62 years old. It is therefore a screen-detected cancer type ($CANTYP = 1$), whose aggressiveness in this case is at a low level according to the rescaled Gleason pattern ($GLE2 = 1$).

After this history profile, it can be determined that the follow-up period of the subject 275 extended from his first visit date, at the age of 58.62 years, to date of his prostate cancer diagnosis when he was 62.62 years old. Hence, this subject represented a complete data due that he experienced the event of interest within the study.

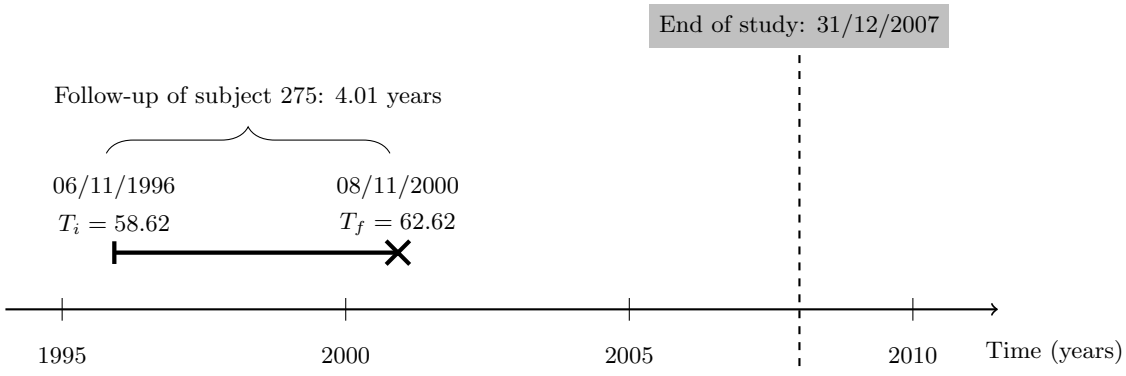

Figure A.3. Subject 275 folow-up within the observation window.

Appendix B

R CODE FOR COMPUTATIONAL ANALYSES

B.1 Code for longitudinal analysis

```
# ============================================================ #
# ===============  LONGITUDINAL DATA ANALYSIS  =============== #
# ============================================================ #

# LOAD THE LIBRARIES
# ==================

library(nlme)
library(Hmisc)

# From the PCa Dataset information, the following variables are selected for each
# subject visits: OBS, ID, AGE0, LLPSA0, START, PSA, LLPSA, STOP, EVENT

# Read de subset of dimensions 4673x9 as a data frame:
lmm_data <- read.table('cox_td_llpsa.txt',header=T)

# Factorization of identifier code:
lmm_data$ID <- as.factor(lmm_data$ID)

# Translation:
lmm_data <- transform(lmm_data, AGE0=AGE0-45, START=START-45, STOP=STOP-45)

# Ghange the name for the further adaption to the JM package
colnames(lmm_data)[5] <- 'OBSTIME'

## MODEL WITH RANDOM INTERCEPT
## ===========================

lmm.1 <- lme(LLPSA~OBSTIME,random=~1|ID, data=lmm_data,
control=lmeControl(msMaxIter=5000, opt="nlminb",niterEM=5000))
summary(lmm.1)
summary(lmm.1)$tTable
lmm.1$logLik  #  1160.039

# Fixed coefficients:
fixef(lmm.1)
# Random effects:
ranef(lmm.1)
# Global coefficients:
coef(lmm.1)
```

```
## MODEL WITH RANDOM SLOPE
## ======================

lmm.2 <- lme(LLPSA~OBSTIME,random=~-1+OBSTIME|ID, data=lmm_data,
control=lmeControl(msMaxIter=5000, opt="nlminb",niterEM=5000))
summary(lmm.2)
summary(lmm.2)$tTable
lmm.2$logLik  #  997.711

# Fixed coefficients:
fixef(lmm.2)
# Random effects:
ranef(lmm.2)
# Global coefficients:
coef(lmm.2)

# Comparison:
anova(lmm.1,lmm.2) # lmm.1

## MODEL WITH RANDOM INTERCEPT AND RANDOM SLOPE
## ===========================================

lmm.3 <- lme(LLPSA~OBSTIME,random=~OBSTIME|ID, data=lmm_data,
control=lmeControl(msMaxIter=5000, opt="nlminb",niterEM=5000))
summary(lmm.3)
summary(lmm.3)$tTable
lmm.3$logLik  #   1214.29

# Fixed coefficients:
fixef(lmm.3)
# Random effects:
ranef(lmm.3)
# Global coefficients:
coef(lmm.3) infividu:

# Comparison
anova(lmm.1,lmm.3)  # lmm.3
anova(lmm.2,lmm.3)  # lmm.3
```

B.2 Code for survival analysis

```
# ========================================================== #
# ================  SURVIVAL DATA ANALYSIS  ================ #
# ========================================================== #

# LOAD THE LIBRARIES
```

```
# ==================

library(survival)
library(Hmisc)

# From the PCa Dataset information, the following variables are selected
# at first visit of each subject: ID, AGE0, PSA0, LLPSA0, TIME, CENS

# Read de subset of dimensions 2415x6 as a data frame:
cox_llpsa0 <- read.table('cox_llpsa0.txt',header=T)

# Translations of time variables: AGE0 and TIME:
cox_llpsa0$AGE0 <- cox_llpsa0$AGE0-45
cox_llpsa0$TIME <- cox_llpsa0$TIME-45

# NON PARAMETRIC SURVIVAL ANALYSIS WITH ALL SUBJECTS
# ================================================

# Data distribution
all <- survfit(Surv(TIME,CENS)~1,data=cox_llpsa0)
summary(all) # 116 diagnosed

# Kaplan-Meier Curve
par(oma=c(0.5,1,0,0),mar=c(4.5,4.5,3,2),las=1,
cex.main=0.9,cex.lab=0.9,cex.axis=0.9)
plot(all,xlim=c(0,40),conf.int=F,mark.time=T,
lty=1,col=1,lwd=2, lab=c(10,10,7), xaxt='n',
main='Kaplan-Meier survival curve for the PCA Dataset', xlab='Age (years)',
ylab=expression(paste(hat(S),"(t): Proportion without prostate cancer diagnosis")))
axis(1,at=c(0,5,10,15,20,25,30,35,40),labels=c(45,50,55,60,65,70,75,80,85))
segments(0,0.9137,29.8,0.9137, lwd=1, lty=2,col=1)
text(7,0.87,"0.914 (74.8 years)",col=1,cex=0.9)
legend("bottomright", y.intersp = 1.15,
title=expression(underline("Total number of subjects: 2415")),
c("Prostate cancer diagnosis: 116 (4.8%)", "Right-censored data: 2299 (95.2%)"),
inset=0.025, bg='gray95',cex=0.9, bty='white' )

# NON PARAMETRIC SURVIVAL ANALYSIS STRATIFYNG BY PSA0 CUT-OFF
# ==========================================================

# New dichotomous variable, named GROUP:
cox_llpsa0$GROUP=ifelse(cox_llpsa0$PSA0>=3,1,0)
sum(cox_llpsa0$PSA0<3)   # 2151 (89.07%)
sum(cox_llpsa0$PSA0>=3)  # 264  (10.93%)

# Kaplan Meier curves by GROUP
pdf('tercils_AGE0.pdf',width=7,height=5)

par(oma=c(0.5,1,0,0), mar=c(4.5,4.5,3,2), las=1,
cex.main=0.9, cex.lab=0.9, cex.axis=0.9)
plot(survfit(Surv(TIME,CENS)~GROUP,cox_llpsa0),xlim=c(0,40),
```

```
col=1:2,conf.int=F,mark.time=F, lwd=1.5, xaxt='n',
main=bquote(bold(paste("Kaplan-Meier survival curves by ",
bolditalic(PSA),"0 cut-off"))), xlab='Age (years) at first visit',
ylab=expression(paste(hat(S),"(t): Proportion without prostate cancer diagnosis")))
summary.G0 <- summary(survfit(Surv(TIME,CENS)~GROUP,subset(cox_llpsa0,GROUP==0)))
summary.G1 <- summary(survfit(Surv(TIME,CENS)~GROUP,subset(cox_llpsa0,GROUP==1)))

lines(c(summary.G0$time,36.7),c(summary.G0$lower,0.927),col=1,lty=2)
lines(c(summary.G0$time,36.7),c(summary.G0$upper,0.965),col=1,lty=2)
lines(c(summary.G1$time,36.24),c(summary.G1$lower,0.643),col=2,lty=2)
lines(c(summary.G1$time,36.24),c(summary.G1$upper,0.770),col=2,lty=2)

axis(1,at=c(0,5,10,15,20,25,30,35,40),labels=c(45,50,55,60,65,70,75,80,85))
text(5,0.90,"0.946 (74.80 years)",col=1,cex=0.8)
text(5,0.66,"0.704 (71.18 years)",col=2,cex=0.8)

segments(0,0.9459,36.7,0.9459,lty=3,lwd=1,col=1)
segments(0,0.7039,26.2,0.7039,lty=3,lwd=1,col=2)
rug(0.9459, ticksize = -0.02, side = 2,col=1)
rug(0.7039, ticksize = -0.02, side = 2,col=2)

dev.off()

# Log-rank test by PSA0 cur-off:
sf <- survdiff(Surv(TIME,CENS)~GROUP,cox_llpsa0,rho=0)
print(sf)

# NON PARAMETRIC SURVIVAL ANALYSIS STRATIFYNG BY AGE0 TERTILES
# ============================================================

# Kaplan Meaier curves by AGE0 tertile
cox_llpsa0$AGE0t <- with(cox_llpsa0,cut2(cox_llpsa0$AGE0,c(10,15)))
levels(cox_llpsa0$AGE0t)[c(1:3)]<-c('Low AGE0','Medium AGE0','High AGE0')

# Kaplan-Meier curves by AGE0 tertile:
pdf('tercils_AGE0.pdf',width=6,height=5)

par(oma=c(0.5,1,0,0), mar=c(4.5,4.5,3,2), las=1,
cex.main=0.9, cex.lab=0.9, cex.axis=0.9)
plot(survfit(Surv(TIME,EVENT)~AGE0t,cox_llpsa0), xlim=c(0,40),
conf.int=F, mark.time=F, xaxt='n',
col=1:4,lwd=c(2,2,2),lab=c(10,10,7),
main=bquote(bold(paste("Kaplan-Meier survival curves by ",
bolditalic(AGE), "0 tertiles"))), xlab='Age (years)',
ylab=expression(paste(hat(S),"(t): Proportion without prostate cancer diagnosis")))
axis(1,at=c(0,5,10,15,20,25,30,35,40),labels=c(45,50,55,60,65,70,75,80,85))
legend("bottomright",y.intersp = 1.2,levels(cox_llpsa0$AGE0t),col=1:3,lty=1,lwd=2,
inset = 0.01, bg='gray95',cex=0.9)

dev.off()
```

```
# SEMI-PARAMETRIC SURVIVAL ANALYSIS WITH ALL SUBJECTS
# ===================================================

# PH Cox model with the baseline covariate LLPSA0

cox1.llpsa0 <- coxph(Surv(TIME,CENS)~LLPSA0,cox_llpsa0)
summary(cox1.llpsa0) # p-value < 0.0001
cox1.llpsa0$loglik   # -758.3271

# PH Cox model with the baseline covariate AGE0xLLPSA0

cox2.llpsa0 <- coxph(Surv(TIME,CENS)~AGE0*LLPSA0,cox_llpsa0)
summary(cox2.llpsa0) # p-value < 0.0001
cox2.llpsa0$loglik   # -774.429
```

B.3 Code for joint model analysis

```
# ========================================================== #
# ===============  JOINT MODEL ANALYSIS  =================== #
# ========================================================== #

# LOAD THE LIBRARIES
# ==================
library(nlme)
library(survival)
library(JM)
library(plotrix)

# From the PCa Dataset information, the following variables are selected at first
# visit of each subject:
# OBS, ID, AGE0, LLPSA0, OBSTIME, LLPSA, STOP, TIME, EVENT, CENS

# Read de subset of dimensions 4673x10 as a data frame:
jm_llpsa <- read.table('jm_llpsa.txt',header=T)

# Change name of START avriable to OBSTIME variable:
colnames(jm_llpsa)[5]<-'OBSTIME'

# Translations:
jm_llpsa <- transform(jm_llpsa, AGE0=AGE0-45,
OBSTIME=OBSTIME-45, STOP=STOP-45, TIME=TIME-45)

# First observations from each subject:
jm_llpsa.id <- jm_llpsa[!duplicated(jm_llpsa$ID),]

#  SURVIVAL MODEL WITH AGE0xLLPSA0
coxfit.id <- coxph(Surv(TIME,CENS)~AGE0:LLPSA0,data=jm_llpsa.id,x=TRUE)

# LONGITUDINAL MODEL WITH RANDOM INTERCEPT
```

```
# ======================================
lmm.1 <- lme(LLPSA~OBSTIME,random=~1|ID, data=jm_llpsa,
control=lmeControl(msMaxIter=5000, opt="nlminb",niterEM=5000))

# LONGITUDINAL MODEL WITH RANDOM SLOPE
# ====================================
lmm.2 <- lme(LLPSA~OBSTIME,random=~-1+OBSTIME|ID, data=jm_llpsa,
control=lmeControl(msMaxIter=5000, opt="nlminb",niterEM=5000))

# LONGITUDINAL MODEL WITH RANDOM INTERCEPT AND SLOPE, UNSTRUCTURED D MATRIX
# ========================================================================
lmm.3 <- lme(LLPSA~OBSTIME,random=~OBSTIME|ID, data=jm_llpsa,
control=lmeControl(msMaxIter=5000, opt="nlminb",niterEM=5000))

# JOINT MODELS WITH LLPSA0xAGE0 SURVIVAL BASELINE COVARIATE
# ========================================================
jm.int.intercept <- jointModel(lmm.1 ,coxfit.id, timeVar="OBSTIME",
method="piecewise-PH-aGH",control = list(iter.EM=5000,optimizer='nlminb'))
summary(jm.int.intercept)
confint(jm.int.intercept)

jm.int.slope <- jointModel(lmm.2, coxfit.id, timeVar="OBSTIME",
method="piecewise-PH-aGH",control = list(iter.EM=5000,optimizer='nlminb'))
summary(jm.int.slope)
confint(jm.int.slope)

jm.int.int_slo <- jointModel(lmm.3, coxfit.id, timeVar="OBSTIME",
method="piecewise-PH-aGH",control = list(iter.EM=5000, optimizer='nlminb'))
summary(jm.int.int_slo)
confint(jm.int.int_slo)

# VALIDATION OF THE FITTED JOINT MODEL: RESIDUAL PLOTS
# ====================================================

pdf('residuals_TFM.pdf',width=7,height=5)

plotResid <- function(x,y,col.loess=1,...){
 plot(x,y,pch=19,col='gray55',cex=0.5,...)
 lines(lowess(x,y), col=col.loess,lwd=3)
 abline(h=0, lty=3, col=1, lwd=1)
}

par(mfrow = c(2,2), oma=c(0,1,0,0), mar=c(4.5,3.75,3,3.25), las=1,
cex.main=0.85, cex.axis=0.85, cex.lab=0.85)

resSubY <- residuals(jm.int.intercept,
process = "Longitudinal",type = "stand-Subject")
fitSubY <- fitted(jm.int.intercept, process = "Longitudinal",
type = "Subject")
plotResid(fitSubY, resSubY, xlab = "Fitted values", ylab ="Residuals",
xlim=c(0,1.5),xaxt='n',ylim=c(-4,4),
```

```
main = "Standardized subject-specific residuals vs fitted values")
axis(1,at=c(0,0.3,0.6,0.9,1.2,1.5),label=c('0.0',0.3,0.6,0.9,1.2,1.5))

resMargY <- residuals(jm.int.intercept,
 process = "Longitudinal",type = "stand-Marginal")
fitMargY <- fitted(jm.int.intercept, process = "Longitudinal",
type = "Marginal")
plotResid(fitMargY, resMargY, xlab = "Fitted values", ylab =
"Residuals",xlim=c(0.3,0.8),xaxt='n',ylim=c(-6,6),yaxt='n',
main = "Standardized marginal residuals vs fitted values")
axis(1,at=c(0.3,0.4,0.5,0.6,0.7,0.8),label=c(0.3,0.4,0.5,0.6,0.7,0.8))
axis(2,at=c(-6,-3,0,3,6),label=c('-6.0','-3.0','0.0','3.0','6.0'))

resMartT <- residuals(jm.int.intercept,
process = "Event", type = "Martingale")
fitSubY <- fitted(jm.int.intercept, process = "Longitudinal",
type = "EventTime")
plotResid(fitSubY, resMartT, xlab = "Fitted values", ylab = "Residuals",
xlim=c(0,1.5),xaxt='n',ylim=c(-1.5,1.5),yaxt='n',
main = "Martingale residuals vs fitted values")
axis(1,at=c(0,0.3,0.6,0.9,1.2,1.5),label=c('0.0',0.3,0.6,0.9,1.2,1.5))
axis(2,at=c(-1.5,0,1.5),label=c('-1.5','0.0','1.5'))

resCST <- residuals(jm.int.intercept,
process = "Event", type = "CoxSnell")
sfit <- survfit(Surv(resCST, CENS) ~ 1, data = jm_llpsa.id)
plot(sfit, mark.time = FALSE, conf.int = TRUE, lty = 1:2,
xlab = "Cox-Snell Residuals", ylab = "Survival probability",
xlim=c(0,1.4),xaxt='n', main = "Survival function of Cox-Snell Residuals")
curve(exp(-x), from=0, to=max(jm_llpsa.id$TIME), add=TRUE, col=1, lwd=2)
axis(1,at=c(0,0.2,0.4,0.6,0.8,1,1.2,1.4),label=c('0.0',0.2,0.4,0.6,0.8,1.0,1.2,1.4))

dev.off()

# SURVIVAL DYNAMICAL PREDICTIONS
# ==============================

# Change the name of variable ID by id:
colnames(jm_llpsa)[2]<-'id'

# The four registres for subject ID=id=845:
jm_llpsa[jm_llpsa$id==845,c(2,4:9)]

set.seed(123) # We set the seed for reproducibility
survPrbs <- survfitJM(jm.int.intercept,
newdata=jm_llpsa[jm_llpsa$id==845, ], M=200)


# The plot method depicts the estimates of the conditional survival probabilities.

pdf('survival_556.pdf',width=7,height=5)
```

```
par(oma = c(0, 1, 0, 1),cex.main=1, cex.lab=1, cex.axis=1)
inicio <- 15
plot.survfitJM(survPrbs,lty = c(1:2,3,3),
main=expression(paste('Conditional survival probabilities for subject ',
italic(i),'=556')), xlim=c(inicio,40), col=rep('gray55',4), lwd=c(2,2,2,2),
conf.int = TRUE, xaxt='n',yaxt='n')
axis(1, at=c(inicio,25,30,35,40), label=c('0','70','75','80','85'))
axis.break(1, 19.5, style="zigzag", brw=0.05)
axis(1, at=c(27.5,32.5,37.5), label=c('','',''),tck=-0.015)
axis(2, at=c(0,0.2,0.4,0.6,0.8,1), label=c('0.0',0.2,'0.4','0.6','0.8','1.0'),las=1)
segments(24.58, 0, x1 = 24.58, y1 = 1, col=1, lty=3, lwd=1)
mtext('Age (years)',las=0,side=1,line = 2.75, cex=0.9)
text(21.6,0.5,'t = 69.58 years',cex=1)

dev.off()
```